ADAPTIVE BASIS SAMPLING FOR SMOOTHING SPLINES

A Dissertation

by

NAN ZHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jianhua Huang
Committee Members,	Raymond J. Carroll
	Mohsen Pourahmadi
	Joel Zinn
Head of Department,	Valen E. Johnson

August 2015

Major Subject: Statistics

ABSTRACT

Smoothing splines provide flexible nonparametric regression estimators. Penalized likelihood method is adopted when responses are from exponential families and multivariate models are constructed with certain analysis of variance decomposition. However, the high computational cost of smoothing splines for large data sets has hindered their wide application. We develop a new method, named adaptive basis sampling, for efficient computation of smoothing splines in super-large samples. Generally, a smoothing spline for a regression problem with sample size n can be expressed as a linear combination of n basis functions and its computational complexity is $O(n^3)$. We achieve a more scalable computation in the multivariate case by evaluating the smoothing spline using a smaller set of basis functions, obtained by an adaptive sampling scheme that uses values of the response variable. Our asymptotic analysis shows that smoothing splines computed via adaptive basis sampling converge to the true function at the same rate as full basis smoothing splines. We show that the proposed method outperforms a sampling method that does not use the values of response variable by simulation studies, and apply it to several real data examples.

To my family

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Jianhua Huang, for his guidance and support throughout my doctoral studies. Professor Huang has been a source of constant personal encouragement and intellectual stimulation not only during the research work but over the five years I have been at Texas A&M. It is my honor to be able to work with him.

My thanks also extend to the other members of my committee, Raymond Carroll, Mohsen Pourahmadi and Joel Zinn, for their interest and insightful comments. I would like to thank Ping Ma for initiating the research projects in this work and our collaboration is a fruitful experience.

I want to thank everyone in the Department of Statistics, including faculty, staff, visitors and my fellow graduate students. I feel fortunate to spend my five years in such a friendly and stimulating environment. In particular, I benefited greatly from many conversations with Anirban and interactions with members in our research group. I am also thankful to my friends outside the department, especially Hancheng, Fan, Bo and Gang. Life out of work is much more colorful with them.

Finally, I dedicate this work to my family: my parents, Dacheng Zhang and Yuexia Sun, and my fiancée, Ruoxin Wang, for their sacrifices and devotion. Their unconditional love and endless support have always been my strength.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Smoothing splines provide flexible nonparametric regression estimators. Due to several distinguished features, smoothing splines stand out as a popular choice among nonparametric modeling methods (Ruppert et al., 2003). First, it is conceptually simple. Smoothing spline is essentially a set of segmented low degree polynomials connected smoothly at some specified points. Second, its model-fitting is data-driven and does not require manual parameter-tuning. Smoothing spline treats each data point as a node and uses a penalized likelihood to fit the model with regularization parameters tuned by generalized cross-validation (Gu, 2013). Nevertheless, the high computational cost of smoothing splines for large data sets has hindered their wide application. For example, modern biology technologies can sequence tens of millions DNA/cDNA fragments in parallel. After the resulting sequences are mapped to genome, one gets a sequence of short read counts along genome. Smoothing splines have been used extensively for modeling and processing single sequencing sample. However, nonparametric joint modeling of multiple sequencing samples are still lacking due to expensive computational cost.

In this dissertation, we develop a new method, named adaptive basis sampling, for efficient computation of smoothing splines in large samples. Such basis sampling scheme makes use of information from response variable and is computationally effective. We consider nonparametric regression with Gaussian and non-Gaussian responses, and present a systematic treatment to analyze the asymptotic properties of the smoothing spline estimator with adaptively selected basis functions.

1.1 Background

To estimate a function of interest η on a generic domain $\mathcal{X}$ using stochastic data, one may use the minimizer of the penalized likelihood

$$L(\eta|\text{data}) + \lambda J(\eta), \tag{1.1}$$

where $L(\eta|\text{data})$ is usually taken as negative log likelihood of the data and $J(\eta)$ is a quadratic functional quantifying the roughness of η. The penalty parameter λ controls the trade-off between the goodness-of-fit and smoothness of η. See Wahba (1990), Gu (2013) and Wang (2011) for overviews of this method.

The standard formulation of smoothing splines performs the minimization of (1.1) in a reproducing kernel Hilbert space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$, where $J(\cdot)$ is seen as a squared semi-norm. Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$ and assume that $\mathcal{N}_J$ is a finite dimensional linear subspace of $\mathcal{H}$. Denote by $\mathcal{H}_J$ the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. The reproducing kernel Hilbert space provides a very general framework for nonparametric regression where the penalty term $J(\eta)$ can be chosen to serve different purposes. For univariate function estimation on a compact interval $\mathcal{X}$, one can use

$$J(\eta) = \int_{\mathcal{X}} (\eta^{(m)})^2 dx.$$

In particular, $m = 2$ corresponds to the commonly-used second derivative penalty and the minimizer of (1.1) is a natural cubic spline. For estimating a multivariate function on a compact domain $\mathcal{X} \subset \mathbb{R}^d (d > 1)$, one can use the thin-plate spline

penalty

$$J_{md}(\eta) = \int \cdots \int_{\mathcal{X}} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1! \cdots \nu_d!} \Big(\frac{\partial^m \eta}{\partial x_1^{\nu_1} \cdots \partial x_d^{\nu_d}} \Big)^2 dx_1 \cdots dx_d \tag{1.2}$$

where m is the order of derivatives and d is the number of predictor variables (Duchon, 1977). As a special case, when $m = 2$ and $d = 2$ we have

$$J_{22}(\eta) = \iint_{\mathcal{X}} \Big(\frac{\partial^2 \eta}{\partial x_1^2} \Big)^2 + \Big(\frac{\partial^2 \eta}{\partial x_1 \partial x_2} \Big)^2 + \Big(\frac{\partial^2 \eta}{\partial x_2^2} \Big)^2 dx_1 dx_2.$$

See Gu (2013) for details about defining the penalty term and corresponding reproducing kernel Hilbert space for modeling a multivariate regression function using smoothing spline analysis of variance models.

Univariate smoothing splines can be computed in $O(n)$ by applying the Reinsch (1967) algorithm. In general, as we shall see in the next chapter, the computational cost of finding the minimizer of (1.1) is in the order of $O(n^3)$ and thus is very large for big data sets. To lower the computational cost, over the past decades, there have been efforts to find sparse sets of basis functions to approximate the minimizer of (1.1). Luo and Wahba (1997) and Zhang et al. (2004) applied variable selection techniques, but it is not clear whether the resulting estimators share the good asymptotic properties of standard smoothing splines. Gu and Kim (2002) and Kim and Gu (2004) developed a simple random sampling approach for basis function selection and established a coherent theory for the convergence of their approximated smoothing splines. To overcome the computational burden of smoothing splines, pseudosplines (Hastie, 1996) and penalized splines (Ruppert et al., 2003) have also been proposed. Both use a small number of fixed basis functions to approximate the smoothing splines; they are similar in spirit to Gu and Kim (2002) and Kim and Gu

(2004) but differ in the construction of the basis functions.

1.2 Main contribution

Our adaptive basis sampling method for approximating smoothing splines is an extension of the simple random sampling approach of Gu and Kim (2002) and Kim and Gu (2004). Its novelty is that we select the basis functions according to the slicing along the range of the response variable. These adaptively selected basis functions form a reduced model space, called the effective model space. We compute the approximated smoothing spline estimator in the reduced space to achieve efficient computation. This adaptive sampling strategy differs from all existing methods based on sampling basis functions on the direction of the predictors. It can recover fine details of the response surface better than the simple random sampling scheme.

With the proposed basis sampling method, we achieve a more scalable computation through a sparse approximation of smoothing spline ANOVA models in a lower dimensional effective model space. The asymptotic analysis shows that smoothing spline estimator computed via adaptive basis sampling converges at the same rate as that of full basis smoothing spline estimator. As evident in our simulation and real data analysis studies, smoothing spline ANOVA models approximation via adaptive basis selection provide very accurate estimates.

Chapter 2 is devoted to nonparametric regression with Gaussian responses. Effective methods for smoothing parameter selection and generic algorithms for computation are the main topics. It is also focusing on the estimation of multivariate functions with large samples. Compared with other competing methods such as penalized spline, our approach is statistically more efficient according to both simulation study and a read data example on geographical imaging analysis. Chapter 3 extends the adaptive basis sampling to a generalized regression problem, where

the response variables come from exponential family distributions. Motivated by two next-generation sequencing data sets, we are particularly interested in modeling counts data and thus responses are assumed to be Poisson, binomial or negative binomial distributed. Chapter 4 studies some theoretical properties of estimators constructed by the adaptive basis sampling method. Proofs of some lemmas and theorems in Chapter 2 and 3 are collected. Since Gaussian distribution is also an exponential family distribution, proofs for Chapter 2 are special cases of those of Chapter 3. We only present the general case.

2. REGRESSION WITH GAUSSIAN RESPONSES

2.1 Introduction

Consider the nonparametric regression model

$$y_i = \eta(x_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{2.1}$$

where y_i is the ith observation of the response variable, x_i is the ith observation of the predictor variable on the domain $\mathcal{X} \subset \mathbb{R}^d$ $(d \geq 1)$, η is the nonparametric function to be estimated, and the ϵ_i's are independent and identically distributed random errors with mean zero and unknown constant variance σ^2. A widely used method for estimating the unknown function η in (2.1) is via minimization of the penalized least squares criterion

$$\text{PLS}(\eta) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta(x_i)\}^2 + \lambda\, J(\eta), \tag{2.2}$$

where $J(\eta)$ is a quadratic functional quantifying the roughness of η. The first term in expression (2.2) discourages lack of fit, and the second term penalizes the roughness of η. The penalty parameter λ controls the trade-off between the goodness-of-fit and smoothness of η. Multivariate penalty parameters can be introduced when estimating a multivariate function, but we focus our presentation on the single penalty case. See Wahba (1990), Gu (2013) and Wang (2011) for overviews of this method, including how to introduce multivariate penalty parameters.

The standard formulation of smoothing splines performs the minimization of (2.2) in a reproducing kernel Hilbert space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$, where $J(\cdot)$ is a squared semi-norm. Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$ and assume that

$\mathcal{N}_J$ is a finite dimensional linear subspace of $\mathcal{H}$ with basis $\{\xi_i : i = 1, \ldots, m\}$, where $m = \dim(\mathcal{N}_J)$. Denote by $\mathcal{H}_J$ the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. Let P be the orthogonal projection operator from $\mathcal{H}$ onto $\mathcal{H}_J$. Then $J(\cdot)$ is a well-defined squared norm of $\mathcal{H}_J$ and for any $\eta \in \mathcal{H}$, $J(\eta) = J(P\eta) = \|P\eta\|^2_{\mathcal{H}_J}$. With this norm, $\mathcal{H}_J$ is also a reproducing kernel Hilbert space, and we denote its reproducing kernel by $R_J(\cdot, \cdot)$.

In this chapter, we propose an adaptive basis sampling method for approximating smoothing splines. We select the basis functions according to the response variable. These basis functions form an effective model space. Efficient computation is achieved when we compute the approximated smoothing spline estimator in the effective model space. In addition, we develop an asymptotic theory on the rate of convergence of our approximated smoothing spline estimator. This theory is non-standard because of the response-dependent sampling scheme, and yields conditions on the dimension of the effective model space to warrant the same convergence rate as the regular smoothing spline estimators. Such conditions provide useful practical guidelines for the sample size of the adaptive sampling.

2.2 Smoothing splines and computational issues

We first state the so-called representer theorem (e.g., Wahba, 1990), which declares that although the original penalized least squares problem for smoothing splines is formulated in the infinite-dimensional function space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$, the solution lies in a finite-dimensional space. Recall that $\mathcal{H}$ has the tensor-sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, $\{\xi_i\}_{i=1}^m$ spans the null space $\mathcal{N}_J$ of the quadratic functional J, and $R_J(\cdot, \cdot)$ is the reproducing kernel of $\mathcal{H}_J$.

Theorem 2.2.1. *There exist vectors* $d = (d_1, \ldots, d_m)^\top \in \mathbb{R}^m$ *and* $c = (c_1, \ldots, c_n)^\top \in$

$\mathbb{R}^n$ *such that the minimizer of* (2.2) *over* $\mathcal{H}$ *can be represented as*

$$\eta(x) = \sum_{k=1}^{m} d_k \xi_k(x) + \sum_{i=1}^{n} c_i R_J(x_i, x), \qquad x \in \mathcal{X}. \tag{2.3}$$

Theorem 2.2.1 implies that we need only search for the minimizer of (2.2) over the collection of functions of form (2.3), so the problem reduces to finding the coefficient vectors d and c that satisfy a system of linear equations. Let $x = (x_1, \ldots, x_n)^\top$ be the vector of observed values of the predictor variable, and $y = (y_1, \ldots, y_n)^\top$ be the vector of corresponding observations of the response variable. Let $\eta = \{\eta(x_1), \ldots, \eta(x_n)\}^\top$ denote the n evaluations of $\eta(\cdot)$ at x, S denote the $n \times m$ matrix with the (i, j)th entry $\xi_j(x_i)$, and R denote the $n \times n$ matrix with the (i, j)th entry $R_J(x_i, x_j)$. Then the decomposition (2.3) applied to x yields the system of equations

$$\eta = Sd + Rc, \tag{2.4}$$

and thus the first term on the right-hand side of (2.2) becomes

$$n^{-1}(y - Sd - Rc)^\top (y - Sd - Rc). \tag{2.5}$$

On the other hand, for any function η with the expansion (2.3), the penalty function $J(\eta)$ in (2.2) can also be written in a matrix form using the reproducing property of $R_J(\cdot, \cdot)$, i.e.,

$$\langle R_J(x_i, \cdot), R_J(x_j, \cdot) \rangle_{\mathcal{H}_J} = R_J(x_i, x_j).$$

Recall that $P : \mathcal{H} \to \mathcal{H}_J$ is a projection operator. For any η as in (2.3), $P\eta =$

$\sum_{i=1}^{n} c_i R_J(x_i, \cdot)$. Hence

$$\begin{aligned} J(\eta) = \|P\eta\|^2_{\mathcal{H}_J} &= \Big\langle \sum_{i=1}^{n} c_i R_J(x_i, x), \sum_{i=1}^{n} c_i R_J(x_i, x) \Big\rangle_{\mathcal{H}_J} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i R_J(x_i, x_j) c_j = c^\top R\, c. \end{aligned} \tag{2.6}$$

Combining (2.5) and (2.6), we see that the penalized least squares criterion (2.2) is reduced to

$$\mathrm{PLS}(\eta) = \frac{1}{n}(y - Sd - Rc)^\top (y - Sd - Rc) + \lambda\, c^\top R\, c. \tag{2.7}$$

Since $\mathrm{PLS}(\eta)$ is a quadratic form in both d and c, its minimizer has a closed-form expression. Differentiating (2.7) with respect to d and c and setting the derivatives to zero, we obtain the linear system of equations

$$\begin{pmatrix} S^\top S & S^\top R \\ R^\top S & R^\top R + n\lambda R \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} S^\top y \\ R^\top y \end{pmatrix}. \tag{2.8}$$

To solve this system, of size $m + n$, the computational cost is generally of the order $O(n^3)$, which can be prohibitive when the sample size n is large. From Theorem 2.2.1, the number of basis functions used to represent the solution is $m + n$, which grows with n. While the m basis functions for $\mathcal{N}_J$ are needed, it may be not necessary to use all n basis functions for $\mathcal{H}_J$. If a smaller number of basis functions can provide a good approximation of the smoothing spline solution, then a computationally efficient algorithm can be developed to handle cases with large sample size. We discuss two sampling approaches for selecting basis functions in the next section.

2.3 Sampling of basis functions

2.3.1 Uniform sampling of basis functions

We first review an approach of selecting basis functions by randomly sampling the observations of the predictor variable and discuss its limitations, and then present our new sampling approach that involves the response variable.

From the representer theorem, each of the n basis functions for representing the function in $\mathcal{H}_J$ is uniquely associated with an observed value of the predictor variable. Thus a natural idea for selecting the basis functions is through randomly sampling the observed values of the predictor variable. Specifically, we draw a random sample of size n^* from the observed predictor values $\{x_i\}_{i=1}^n$, denoted as $x^* = (x_1^*, \ldots, x_{n^*}^*)^\top$, and use the corresponding basis functions, $\{R_J(x_i^*, x)\}_{i=1}^{n^*}$, to represent functions in $\mathcal{H}_J$. We then solve the penalized least squares problem in the effective model space $\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_i^*, x), i = 1, \ldots, n^*\}$. When n^* is much smaller than n, the computational cost can be significantly reduced.

Gu and Kim (2002) and Kim and Gu (2004) proved that this uniform sampling scheme has some nice theoretical properties. Under some reasonable conditions, the smoothing spline estimator computed under this scheme can achieve the same asymptotic convergence rate as the full basis smoothing spline estimator that uses all the basis functions indicated in the representer theorem.

When the number of sampled basis functions increases, the estimator from the uniform sampling strategy will approach the smoothing spline estimator and reveal the underlying true function. However, if constrained by computational resources, one may not sample enough basis functions to achieve a satisfactory result. Figure 2.1 illustrates this with a toy example. The underlying true function is the density function of a two-component mixture of normal distributions. Panel (c) shows the

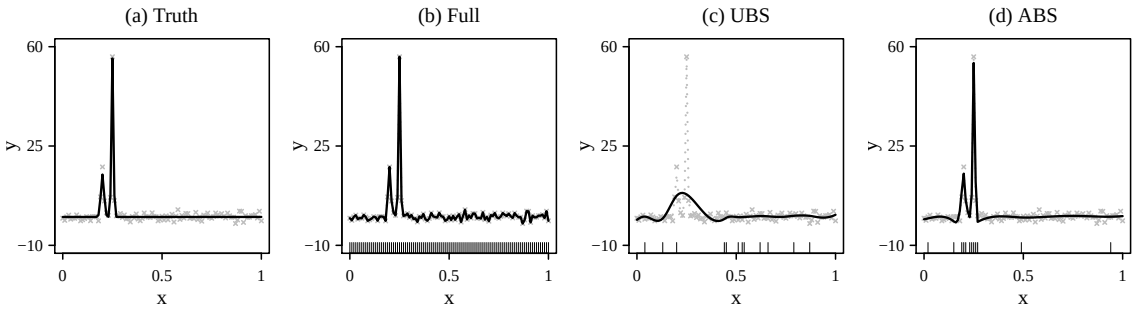

Figure 2.1: Toy regression function with two close peaks: (a) true signal (solid line) and 100 observations (gray crosses); (b) smoothing spline fit (solid line) with full basis; (c) smoothing spline fit (solid line) with 12 uniformly sampled basis functions (UBS); (d) smoothing spline fit (solid line) with 12 adaptively sampled basis functions (ABS). In (b)-(d), short vertical lines at the bottom mark the data points corresponding to the selected basis functions; observations are indicated by gray crosses; true signal is shown as dotted gray lines.

smoothing spline fit using 12 uniformly sampled basis functions, which does not reveal the two peaks of the mixture components because uniform sampling does not select the basis function corresponding to the point with the largest y-value. Unless the number of basis functions is greatly increased, there is little chance that the estimator can capture this peak.

2.3.2 Adaptive sampling of basis functions

We propose a new sampling scheme to select basis functions which makes use of the observed values of the response variable. This scheme may sample more basis functions in regions where the response function has big changes and sample fewer basis functions where the response surface is relatively flat. We call this new scheme adaptive basis sampling.

Like the uniform sampling scheme discussed in §2.3.1, adaptive sampling also

samples the basis functions from the collection $\{R_J(x_i, \cdot) : i = 1, \ldots, n\}$ as indicated in the representer theorem. The difference is the way the sampling is performed. In adaptive basis sampling, we first group the x_i's according to the corresponding value of the response variable, and then draw random samples within each group. The detailed procedure is given below.

1. Divide the range of the responses $\{y_i\}_{i=1}^n$ into K disjoint intervals, S_1, ..., S_K. Let $|S_k|$ denote the number of observations in S_k.

2. For each S_k, consider the collection of all pairs (x_i, y_i) where $y_i \in S_k$, and draw a random sample of size n_k from this collection. Denote the sampled predictor values by $x^{*(k)} = (x_1^{*(k)}, \ldots, x_{n_k}^{*(k)})$.

3. Combine $x^{*(1)}, \ldots, x^{*(K)}$ together to form a set of sampled predictor values $\{x_1^*, \ldots, x_{n^*}^*\}$. This set has size $n^* = \sum_{k=1}^K n_k$.

4. Form the effective model space

$$\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_j^*, \cdot),\ j = 1, \ldots, n^*\}. \tag{2.9}$$

   Minimize the penalized least squares criterion (2.2) over this effective model space.

The first step of the adaptive basis sampling procedure groups together observations with similar response values. It is the same operation as binning when constructing histograms and slicing in sliced inverse regression (Li, 1991; Cook, 1998). Each set $\{(x_i, y_i) : y_i \in S_k\}$ is referred to as a slice of the data. We expect this adaptive sampling scheme to select more effective basis functions than uniform sampling.

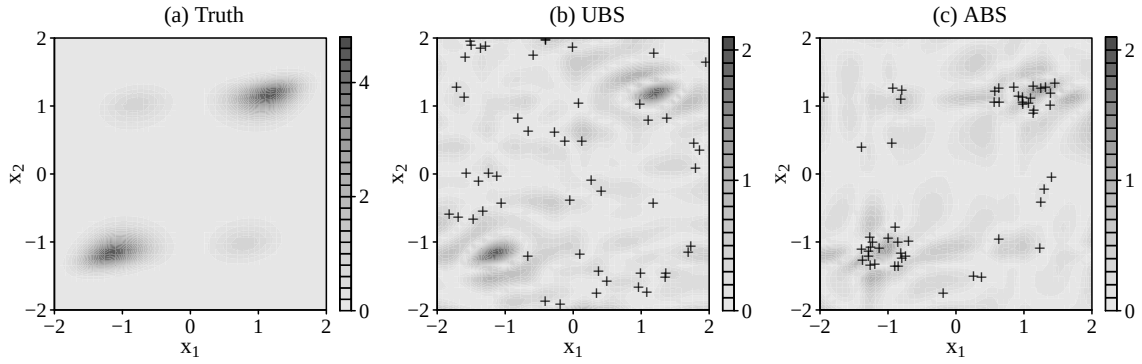

Figure 2.2: Bivariate nonparanormal copula density function. (a): contour plot of true function; (b)–(c): contour plots of absolute values of fitting residuals by smoothing splines based on uniform basis sampling (UBS) and adaptive basis sampling (ABS). The sampled basis functions are marked by +'s.

Figure 2.1(d) displays the smoothing spline fit from the adaptive sampling scheme with 12 basis functions. The fit reveals the two peaks of the mixture components well, since basis functions corresponding to the peak points are sampled.

To further illustrate how adaptive basis sampling works and compare it with uniform basis sampling, we considered a two-dimensional example for which the response surface is a bivariate nonparanormal copula density function; see §2.5 for its analytical form. Figure 2.2(a) depicts the contour plot of the true function, showing four peaks: two are significantly higher than the others. Contour plots of absolute values of residuals after smoothing spline fitting, presented in Fig. 2.2(b)-(c), indicate that the estimated two big peaks from adaptive basis sampling are closer to the truth than from uniform basis sampling. That adaptive basis sampling smoothing spline yields a better estimate can be explained by the distribution of

sampled basis functions, also shown in (b) and (c): the basis functions sampled by uniform basis sampling spread over the whole domain while those sampled by adaptive basis sampling are mainly distributed around the four peaks, especially the two significant ones.

In §2.4, we show that the adaptive sampling scheme can achieve the asymptotic rate of convergence of the original smoothing spline estimator, although a much smaller set of basis functions is employed. The theoretical results of Gu and Kim (2002) and Kim and Gu (2004) for uniform sampling cannot be applied to adaptive sampling, because values of the response variable are used in selecting the basis functions.

2.3.3 Efficient computation

We now present the details of the computational algorithm when adaptive basis sampling is used to compute the smoothing spline estimator. Recall that the selected data points are denoted by $x^* = (x_1^*, \ldots, x_{n^*}^*)^\top$. Under adaptive basis sampling, the minimizer of (2.2) is approximated by

$$\eta_A(x) = \sum_{k=1}^{m} d_k \xi_k(x) + \sum_{j=1}^{n^*} c_j R_J(x_j^*, x).$$

We let S denote the $n \times m$ matrix with (i,j)th entry $\xi_j(x_i)$. Let R_* be a $n \times n^*$ matrix with the (i,j)th entry $R_J(x_i, x_j^*)$ and R_{**} be a $n^* \times n^*$ matrix with the (i,j)th entry $R_J(x_i^*, x_j^*)$. If we rearrange the original data by putting the selected data points x^* at the front, R_* is just the left part of R while R_{**} is the top-left corner of R. The evaluations of η_A at locations x, $\eta_A = \{\eta_A(x_1), \ldots, \eta_A(x_n)\}^\top$, satisfy

$$\eta_A = S d_A + R_* c_A, \tag{2.10}$$

where $d_A = (d_1, \ldots, d_m)^\top$ and $c_A = (c_1, \ldots, c_{n^*})^\top$.

Similar to (2.7), we have

$$\mathrm{PLS}(\eta_A) = \frac{1}{n}(y - Sd_A - R_* c_A)^\top (y - Sd_A - R_* c_A) + \lambda\, c_A^\top R_{**}\, c_A, \tag{2.11}$$

whose minimizer $(\hat{d}_A, \hat{c}_A)$ satisfies the linear system of equations

$$\begin{pmatrix} S^\top S & S^\top R_* \\ R_*^\top S & R_*^\top R_* + n\lambda R_{**} \end{pmatrix} \begin{pmatrix} d_A \\ c_A \end{pmatrix} = \begin{pmatrix} S^\top y \\ R_*^\top y \end{pmatrix}. \tag{2.12}$$

System (2.12) can be solved using a method described in Golub and Van Loan (1989). First, a pivoted Cholesky decomposition is performed such that the first matrix on the left-hand side of (2.12) equals $G^\top G$, where G is an upper triangular matrix. Then, forward and backward substitutions are used to solve the system of equations to obtain the estimated coefficients. However, care should be taken when R_* is singular, i.e., the bottom diagonal elements of G are zeros. Kim and Gu (2004) suggested replacing those zeros by an appropriate small value δ and proceeding as if R_* is of full rank.

A standard method for data-driven choice of the penalty parameter λ is to minimize the generalized cross-validation criterion (Craven and Wahba, 1979). To give a formal definition of this, note that the fitted values $\hat{y} = (\hat{\eta}_A(x_1), \ldots, \hat{\eta}_A(x_n))^\top$ can be obtained from the estimated coefficients as $\hat{y} = S\hat{d}_A + R_*\hat{c}_A$. In light of (2.12), $\hat{y} = A(\lambda)y$, where $A(\lambda)$ is the smoothing matrix

$$A(\lambda) = (S, R_*) \begin{pmatrix} S^\top S & S^\top R_* \\ R_*^\top S & R_*^\top R_* + n\lambda R_{**} \end{pmatrix}^+ \begin{pmatrix} S^\top \\ R_*^\top \end{pmatrix}, \tag{2.13}$$

and C^+ denotes the Moore–Penrose inverse of C. The criterion is defined as

$$\mathrm{GCV}(\lambda) = \frac{n^{-1} y^\top \{I - A(\lambda)\}^2 y}{[n^{-1}\mathrm{tr}\{I - A(\lambda)\}]^2}, \tag{2.14}$$

and we minimize it as a function of the penalty parameter λ (Tenorio et al., 2011), using standard nonlinear optimization algorithms. We use the modified Newton algorithm developed by Dennis and Schnabel (1996).

Now we calculate the computational complexity, using the fact that $m \ll n^* \ll n$ to simplify the expressions. The construction of the linear system (2.12) is of the order $O(nn^{*2})$, the Cholesky decomposition takes $O(n^{*3})$ flops, the subsequent forward and backward substitutions take $O(n^{*2})$ flops respectively, and the evaluation of (2.14) requires the calculation of $\mathrm{tr}\{A(\lambda)\}$, which takes $O(nn^{*2})$ flops. The overall computational cost is of the order $O(nn^{*2})$.

2.3.4 Bayesian confidence intervals

The efficient computational scheme can also be used to compute Bayesian confidence intervals (Wahba, 1983). Bayesian confidence intervals have certain across-the-function coverage property (Nychka, 1988). We need modify Wabha's formulation slightly to take into account the fact that the basis used adaptive sampling is not a full basis.

Analogous to Wahba (1983), we decompose $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in the space $\mathcal{N}_J$ and η_1 has an independent Gaussian process prior with mean zero and covariance

$$\mathrm{E}\{\eta_1(x_k)\eta_1(x_l)\} = \frac{\sigma^2}{n\lambda} R_J(x_k, x^{*T}) R_{**}^{+} R_J(x^*, x_l),$$

where $x^* = (x_1^*, \ldots, x_{n^*}^*)$, and $R_J(x_k, x^{*T})$ and $R_J(x^*, x_l)$ denote respectively the row

and column vectors $R_J(x_k, x^{*T}) = (R_J(x_k, x_1^*), \ldots, R_J(x_k, x_{n^*}^*))$ and $R_J(x^*, x_l) = R_J(x_l, x^{*T})^T$.

With the priors for η specified above, the posterior mean of $\eta(x)$ has the following expression,

$$\mathrm{E}\{\eta(x) \mid y\} = \xi(x)^\top \hat{d}_A + r(x)^\top \hat{c}_A,$$

where $\xi(x) = (\xi_1(x), \ldots, \xi_m(x))^\top$ is a $m \times 1$ vector, $r(x) = R_J(x^*, x)$ is a $n^* \times 1$ vector, and $\hat{d}_A$ and $\hat{c}_A$ are solutions of (13) in previous section. The posterior variance has the following expression

$$\begin{aligned}\frac{n\lambda}{\sigma^2}\mathrm{var}\{\eta(x) \mid y\} &= r(x)^\top R_{**}^+ r(x) + \xi(x)^\top (S^\top W_*^{-1} S)^{-1} \xi(x) \\ &\quad - 2\xi^\top (S^\top W_*^{-1} S)^{-1} S^\top W_*^{-1} R_* R_{**}^+ r(x) \\ &\quad - r(x)^\top R_{**}^+ R_*^\top (W_*^{-1} - W_*^{-1} S (S^\top W_*^{-1} S)^{-1} S^\top W_*^{-1}) R_* R_{**}^+ r(x),\end{aligned}$$

where $W_* = R_* R_{**}^+ R_*^\top + n\lambda I$. Then we construct the $100(1-\alpha)\%$ Bayesian confidence interval as $\mathrm{E}\{\eta(x) \mid y\} \pm \Phi^{-1}(1-\alpha/2)[\mathrm{var}\{\eta(x) \mid y\}]^{1/2}$, where $\Phi^{-1}(1-\alpha/2)$ is the $100(1-\alpha/2)$ percentile of the standard Gaussian distribution.

2.4 Convergence rates for function estimation

2.4.1 Regularity conditions

We first introduce an inner product associated with the marginal density $f_X(\cdot)$ of the predictor variable X. For any g_1 and g_2 in $\mathcal{L}_2(\mathcal{X})$, define

$$V(g_1, g_2) = \langle g_1, g_2 \rangle = \int_{\mathcal{X}} g_1(x) g_2(x) f_X(x) dx.$$

The norm induced by this inner product is a weighted version of the $\mathcal{L}_2$-norm and the weighting function is the marginal density of the predictor. We define the mean

squared error of the estimator $\hat{\eta}_A$ in estimating the regression function η as the quadratic functional

$$V(\hat{\eta}_A - \eta) = \|\hat{\eta}_A - \eta\|^2 = \langle \hat{\eta}_A - \eta, \hat{\eta}_A - \eta \rangle = \int_{\mathcal{X}} \{\hat{\eta}_A(x) - \eta(x)\}^2 f_X(x) dx.$$

This is a common measure in studying statistical properties of smoothing splines (e.g., Gu and Qiu, 1994).

In the literature, the convergence rate of smoothing splines is usually characterized by an eigen-analysis of the penalty functional J with respect to the quadratic functional V. We now state two commonly-used technical conditions (Gu, 2013). A quadratic functional B is said to be completely continuous with respect to another quadratic functional A, if for any $\epsilon > 0$, there exists a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_1(\eta) = \cdots = L_k(\eta) = 0$ implies that $B(\eta) \leqslant \epsilon A(\eta)$; see Weinberger (1974, §3.3).

Condition C.1. The functional V is completely continuous with respect to J.

By Theorem 3.1 of Weinberger (1974), Condition C.1 implies that V and J can be simultaneously diagonalized; see, e.g., Silverman (1982) and Gu (2013, §9.1). More precisely, there exist a sequence of eigenfunctions $\phi_\nu \in \mathcal{H}$ and the associated nonnegative sequence of eigenvalues $\rho_\nu \uparrow \infty$ such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$ where $\delta_{\nu\mu}$ is the Kronecker delta. Furthermore, any function f satisfying $J(f) < \infty$ can be expressed as a Fourier series expansion $f = \sum_\nu f_\nu \phi_\nu$, where $f_\nu = V(f, \phi_\nu)$.

Condition C.2. For some $r > 1$ and $\beta > 0$, $\rho_\nu > \beta \nu^r$ for sufficiently large ν.

This condition on the growth rate of the eigenvalues is essentially a requirement on the smoothness of $\eta \in \mathcal{H}$. For one-dimensional cubic spline smoothing on a compact interval $\mathcal{X}$ with $J(\eta) = \int_{\mathcal{X}} \{\eta''\}^2$, Conditions C.1 and C.2 are satisfied with $r = 4$

when $V(\eta)$ is equivalent to the standard L_2 norm (Utreras, 1981). For thin-plate splines on a bounded domain of $\mathcal{X} \in \mathbb{R}^d$ with the penalty (1.2), Conditions C.1 and C.2 are satisfied with $r = 2m/d$. For tensor product smoothing splines with penalty $J(\eta) = \sum_{\beta=1}^{s} \theta_\beta^{-1} \|P_\beta \eta\|^2_{\mathcal{H}^\beta}$, one can prove that Condition C.1 holds using the same argument in Example 9.2 of Gu (2013), and Condition C.2 holds with $r = 4 - \epsilon$, where $\epsilon > 0$ (Wahba, 1990).

Condition C.3. For a constant $C < \infty$, $\text{var}\{\phi_\nu(X)\phi_\mu(X)\} \leq C$ for all ν and μ.

Since ϕ_ν is an orthonormal system relative to $V(\cdot, \cdot)$, i.e.,

$$V(\phi_\nu, \phi_\mu) = \int_{\mathcal{X}} \phi_\nu(x)\phi_\mu(x) f_X(x) dx = \delta_{\nu\mu},$$

we have that

$$\text{var}\{\phi_\nu(X)\phi_\mu(X)\} = \int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x) f_X(x) dx - \delta_{\nu\mu}.$$

Thus Condition C.3 is equivalent to the requirement that $\int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x) f_X(x) dx$ is uniformly bounded for all ν and μ.

2.4.2 Convergence rates

This section presents our main results on convergence rates. All proofs are given in Chapter 4.

In our adaptive sampling scheme, the search for the smoothing spline estimator is restricted to the effective model space $\mathcal{H}_E$. We first establish a lemma that justifies the use of the effective model space. Let $\mathcal{H} \ominus \mathcal{H}_E$ denote the orthogonal complement of $\mathcal{H}_E$ in the reproducing kernel Hilbert space $\mathcal{H}$.

Lemma 2.4.1. *As $\lambda \to 0$ and $n^* \lambda^{2/r} \to \infty$, if the function h is not in the effective*

model space, i.e., $h \in \mathcal{H} \ominus \mathcal{H}_E$, we have $V(h) = o_p\{\lambda J(h)\}$.

This result suggests that compared to $\lambda J(h)$, $V(h)$ is negligible when h is orthogonal to $\mathcal{H}_E$, and implies that the space orthogonal to the effective model space $\mathcal{H}_E$ is effectively suppressed by the penalty $\lambda J(\eta)$. Hence, we can capture the essential features of the true function η_0 by restricting the estimator to the effective model space $\mathcal{H}_E$.

For completeness, we state below a standard result for the convergence rate of smoothing splines (e.g., Theorem 9.17 of Gu, 2013).

Theorem 2.4.1. *If $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ for some $p \in [1, 2]$, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, then $(V + \lambda J)(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p)$.*

We now present our main result on the convergence rate of smoothing spline estimator based on the proposed adaptive basis sampling scheme.

Theorem 2.4.2. *If $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ for some $p \in [1, 2]$, as $\lambda \to 0$ and $n^*\lambda^{2/r} \to \infty$, then $(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p)$. In particular, when $\lambda \asymp n^{-r/(pr+1)}$, the estimator achieves the optimal convergence rate,*

$$(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p\{n^{-pr/(pr+1)}\}.$$

This theorem states that, under regularity conditions, the convergence rate of the smoothing spline estimator using an adaptively sampled basis equals that of the smoothing spline estimator using the full basis indicated by the representer theorem. The parameter p in the condition yields a faster rate of convergence for certain functions: for the roughest η satisfying $J(\eta) < \infty$, we have $p = 1$, whereas for the smoothest η, we have $p = 2$; see Wahba (1985) for details.

Note that $J(\eta_0) = \sum_i \rho_i V(\eta_0, \phi_i)^2$. When $J(\eta_0) < \infty$, the condition in Theorem 2.4.2 holds with $p = 1$, and the convergence rate is $O_p(n^{-r/(r+1)})$. When η_0 is in the Sobolev space $W^{m,2}$ on a bounded domain in $\mathbb{R}^d$, we have $r = 2m/d$ and theorem yields the convergence rate $n^{-2m/(2m+d)}$, which is the optimal rate of convergence (Stone, 1982). For the case $d = 1$, Claeskens et al. (2009) and Wang et al. (2011) showed that penalized splines can also achieve the optimal rate of convergence.

Theorem 2.4.2 helps determine the dimension of the effective model space $\mathcal{H}_E$. With $\lambda \asymp n^{-r/(pr+1)}$, Lemma 2.4.1 and Theorem 2.4.2 require that $n^*\lambda^{2/r} \to \infty$, which suggests that a suitable choice of n^* should satisfy $n^* \asymp n^{2/(pr+1)+\delta}$, where δ is an arbitrary small positive number. For univariate cubic smoothing splines with the penalty $J(\eta) = \int_0^1 (\eta'')^2$, $r = 4$ and $\lambda \asymp n^{-4/(4p+1)}$, a suitable choice of the dimension of the effective model space is $n^* = n^{2/(4p+1)+\delta}$, which lies in the interval $(O(n^{2/9+\delta}), O(n^{2/5+\delta}))$ for p taking values in $[1, 2]$. For tensor-product splines, $r = 4-\epsilon$, where $\epsilon > 0$, a suitable choice of the dimension of effective model space is $n^* = n^{2/(4p+1)+\delta}$, which is roughly in interval $(O(n^{2/9+\delta}), O(n^{2/5+\delta}))$. In our simulation study and real data analysis, we take the dimension of the effective model space n^* to be between $5n^{2/9}$ and $20n^{2/9}$.

2.5 Simulation results

Using simulated multivariate regression functions, we compared the smoothing spline estimators based on adaptive basis sampling and uniform basis sampling in terms of estimation accuracy and computational time. We also compared adaptive basis sampling with fast bivariate P-splines, an efficient algorithm for bivariate spline smoothing (Xiao et al., 2013).

Some of our simulation setups involve the joint probability density of a p-dimensional

nonparanormal distribution (Liu et al., 2009)

$$\eta_{\mathsf{copula}}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\{f(x)-\mu\}^\top \Sigma^{-1}\{f(x)-\mu\}\right] \prod_{j=1}^{p} |f_j'(x_j)|, \quad (2.15)$$

where $\mu = 0$, Σ has ones as diagonal entries, 0.5 as off-diagonal elements, and

$$f_j(x) = \alpha_j \operatorname{sign}(x)\,|x|^{\alpha_j}, \qquad j = 1, \ldots, p.$$

This is essentially a probability density function for a Gaussian copula model.

We generated data according to model (2.1) where the predictor variable x was randomly generated from the uniform distribution over the domain of interest. The signal-to-noise ratio, defined as $\text{var}\{\eta(X)\}/\sigma^2$, was set to three levels: $10, 2, 0.4$. For each simulation setup, samples of size $n = 1600$ were generated. We considered four regression function settings:

1. a bivariate blocks function, $\eta_{\mathsf{blocks}}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = \mathsf{blocks}(x_{\langle 1\rangle})$, where $\mathsf{blocks}(\,)$ is the univariate blocks function used in (Donoho and Johnstone, 1994). It has frequent and irregular abrupt changes in one direction and stays constant in the other. The domain of interest is the unit square;

2. a bivariate copula function, given in (3.10), with $p = 2$, $\alpha_1 = 2$, $\alpha_2 = 3$. The domain of interest is $[-2, 2]^2$;

3. a 4-d additive function, $\eta(x) = \eta_{\mathsf{blocks}}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) + \eta_{\mathsf{copula}}(x_{\langle 3\rangle}, x_{\langle 4\rangle})$, where η_{blocks} and η_{copula} are as in setups 1 and 2;

4. a 6-d copula function, the function given in (3.10), with $p = 6$ and $\alpha_j = 0.1$ for all j. The domain of interest is $[-1, 1]^6$.

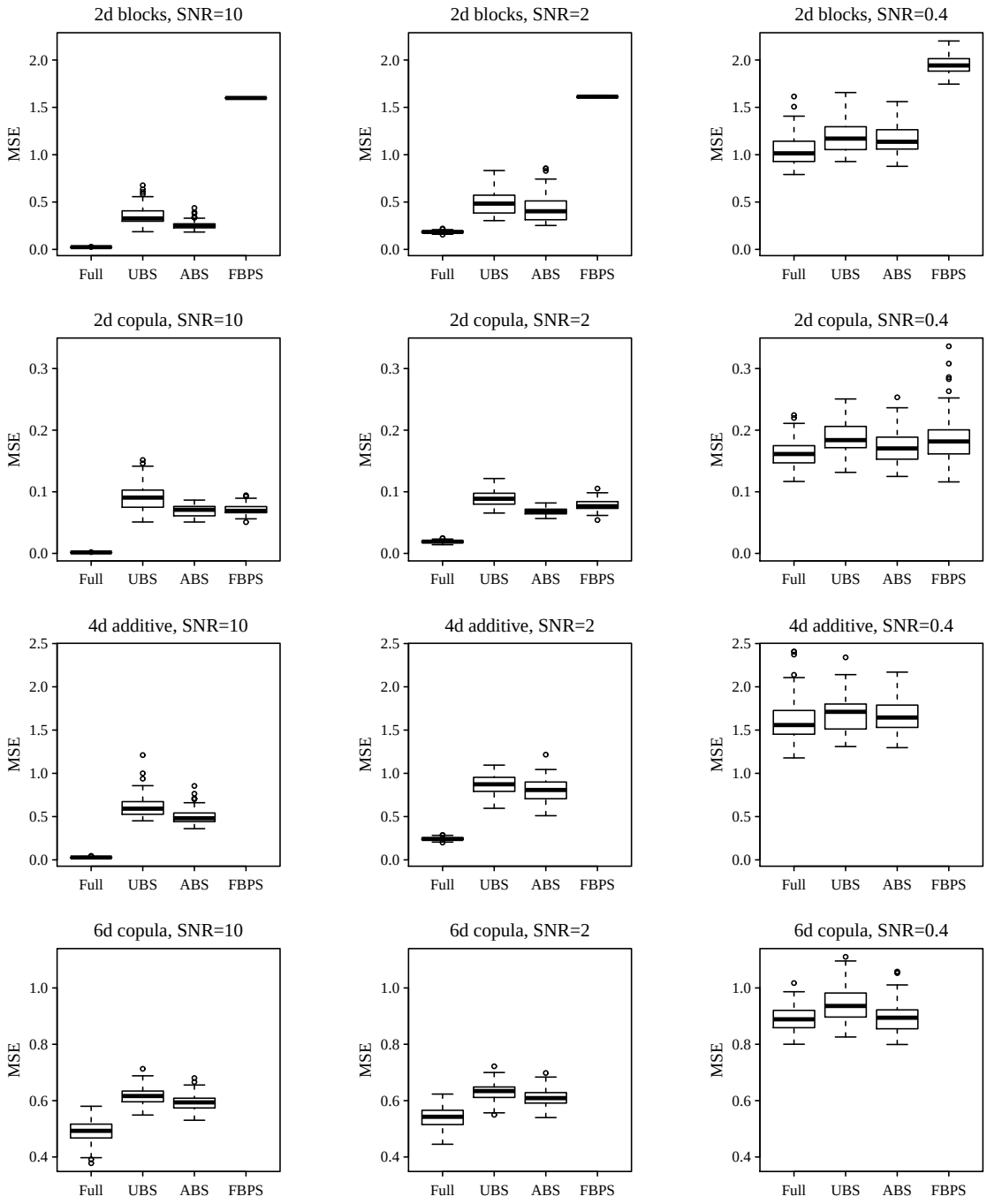

Figure 2.3: Boxplots of the mean squared errors for four multivariate test functions under three signal-to-noise ratios (SNR) $(10, 2, 0.4)$, based on 100 simulation runs. Full, UBS and ABS stand for smoothing spline estimators with full basis, uniform basis sampling and adaptive basis sampling. FBPS is fast bivariate P-splines.

For all four settings, we computed the smoothing spline estimator using the full basis, and using the bases chosen by adaptive basis sampling and uniform basis sampling. For adaptive basis sampling, the number of slices was chosen based on the Scott (1992) method and based on the asymptotic results, the dimension of the effective model space was set to be $10n^{2/9}$, so $n^* = 52$ basis functions were sampled. For a fair comparison, the same number of basis function was used for uniform basis sampling. A thin-plate penalty was used and the penalty parameter λ was selected by minimizing the generalized cross-validation criterion. For cases with dimension higher than two, we assumed a smoothing spline analysis of variance model with second-order interactions to deal with the curse of dimensionality. For the two bivariate setups, we also applied fast bivariate P-splines (Xiao et al., 2013), for which the number of interior knots for each predictor variable was set to be 11, yielding 121 interior knots in total.

To assess the estimation accuracy, we calculated the mean squared error for an estimator, which is defined as $n^{-1}\sum_{i=1}^{n}\{\hat{\eta}(x_i)-\eta(x_i)\}^2$. Figure 2.3 presents boxplots of the mean squared errors based on 100 runs for each setup under three signal-to-noise ratios. For all setups, adaptive basis sampling provides more accurate smoothing spline estimation than uniform basis sampling. Both methods yield higher mean squared errors than the full basis smoothing spline, but this is the price paid for efficient computation with large data sets. When the signal-to-noise ratio decreases, the mean squared error for all methods gets larger and the differences among the methods diminish.

Under the two bivariate settings, adaptive basis sampling performs as well as the fast bivariate P-splines of Xiao et al. (2013) for the bivariate copula function and significantly outperforms it for the bivariate blocks function. The bivariate blocks test function is an extension of the univariate blocks function commonly used to illustrate

univariate spatial adaptive smoothers (Donoho and Johnstone, 1994). However, our proposed method is not designed to achieve spatial adaptivity, since spatial adaptivity requires using location-varying penalty parameters, an idea extensively studied for univariate smoothing splines (Pintore et al., 2006; Liu and Guo, 2010; Wang et al., 2013).

Table 2.1: Means and standard errors (in parentheses) of computational time (in seconds) for four multivariate cases, based on 100 simulation runs. (SNR, signal-to-noise ratio; UBS, uniform basis sampling; ABS, adaptive basis sampling; FBPS, fast bivariate P-splines.)

True function	SNR	Full basis	UBS	ABS	FBPS
2d blocks	10	399 (12)	5.20 (0.12)	5.14 (0.10)	1.38 (0.03)
	2	408 (9)	7.16 (0.35)	6.40 (0.23)	1.41 (0.02)
	0.4	361 (7)	5.00 (0.17)	4.99 (0.17)	1.51 (0.02)
2d copula	10	260 (3)	6.56 (0.20)	6.41 (0.21)	1.63 (0.03)
	2	301 (6)	6.86 (0.18)	6.71 (0.33)	1.59 (0.03)
	0.4	317 (8)	4.69 (0.16)	4.79 (0.14)	1.58 (0.03)
4d blocks+copula	10	1247 (26)	15.16 (0.60)	13.84 (0.59)	–
	2	1222 (25)	16.62 (0.96)	15.54 (0.76)	–
	0.4	1135 (19)	13.16 (0.66)	13.27 (0.60)	–
6d copula	10	9336 (223)	162.88 (7.27)	145.14 (7.32)	–
	2	9572 (283)	179.12 (7.52)	181.27 (6.60)	–
	0.4	7639 (161)	143.10 (6.80)	135.01 (6.81)	–

Table 2.1 summarizes the CPU times of all methods based on 100 runs using Intel Xeon 2.90GHz processor with 64GB of DDR3 RAM. The computing time for the full basis smoothing spline estimator is tens or hundreds times more than that for the basis sampling methods, and for the bivariate cases, the fast bivariate P-spline is the fastest in computation.

2.6 Real data example

At a depth of 2890 km in the Earth, the core-mantle boundary separates turbulent flow of liquid metals in the outer core from slowly convecting, highly viscous mantle silicates. The core-mantle boundary marks the most dramatic change in dynamic processes and material properties in our planet, and accurate images of the structure at or near it over large regions are important for our understanding of the geodynamical processes and the thermo-chemical structure of the mantle and mantle-core system.

To accurately image the core-mantle boundary region, Wang et al. (2006) and Ma et al. (2007) developed a generalized Radon transform to construct raw point images, and applied the smoothing spline method to the raw images. In particular, they extracted seismic waves reflected at core-mantle boundary regions from the public data management center of the Incorporated Research Institutions for Seismology. The seismic waves extracted were generated by around 1300 earthquakes with magnitude mb > 5.2 that occurred between 1988 and 2002, and were recorded at one or more of a total of nearly 1200 stations in central America. Along a 2500 km strip, they then constructed point images of core-mantle boundary regions using a generalized Radon transform. They constructed 163,713 point images at various depths and locations of the strip. At each depth and location, the point images constructed contain many noisy replicates resulting from different reflection angles of the seismic waves, so further statistical analysis is necessary to estimate the true image. In order to be computationally feasible, they estimated the true image using smoothing splines at each location and interpolated the estimated images from all locations to get the three-dimensional image. The image shows peaks of very different magnitudes at several unexpected locations (van der Hilst et al., 2007).

In this section, we apply a smoothing spline with adaptive basis sampling directly to all point images to estimate the three-dimensional image. We let y_{ij} denote the point image at the ith distance, $x_{\langle 1\rangle}$, and the jth depth, $x_{\langle 2\rangle}$. We consider the following model for the point images

$$y_{ij} = \eta(x_{\langle 1\rangle i}, x_{\langle 2\rangle j}) + \epsilon_{ij}.$$

Since the sample size is $n = 163,713$, the regular tensor product smoothing spline is computationally prohibitive. Instead, we apply our cubic tensor product smoothing spline with adaptive basis sampling to the data set with $K = 10$ slices and let the dimension of the effective model space be $n^* = 155$. Define $k_1(u) = u - 0.5$,

$$k_2(x) = \frac{1}{2}\{k_1^2(x) - \frac{1}{12}\}, \qquad k_4(x) = \frac{1}{24}\{k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\},$$

and $R(u_1, u_2) = k_2(u_1)k_2(u_2) - k_4(|u_1 - u_2|)$. The cubic tensor product smoothing spline estimator with adaptive basis sampling has the form

$$\eta(x) = \sum_{\nu=1}^{4} d_\nu \phi_\nu(x) + \sum_{j=1}^{n^*} c_j R_J(x_j^*, x),$$

where $\phi_1(x) = 1$, $\phi_2(x) = k_1(x_{\langle 1\rangle})$, $\phi_3(x) = k_1(x_{\langle 2\rangle})$, $\phi_4(x) = k_1(x_{\langle 1\rangle})k_1(x_{\langle 2\rangle})$ and

$$\begin{aligned} R_J(x, y) =& \theta_1 R(x_{\langle 1\rangle}, y_{\langle 1\rangle}) + \theta_2 R(x_{\langle 2\rangle}, y_{\langle 2\rangle}) \\ &+ \theta_3 R(x_{\langle 1\rangle}, y_{\langle 1\rangle}) k_1(x_{\langle 2\rangle}) k_1(y_{\langle 2\rangle}) + \theta_4 R(x_{\langle 2\rangle}, y_{\langle 2\rangle}) k_1(x_{\langle 1\rangle}) k_1(y_{\langle 1\rangle}) \\ &+ \theta_5 R(x_{\langle 1\rangle}, y_{\langle 1\rangle}) R(x_{\langle 2\rangle}, y_{\langle 2\rangle}). \end{aligned}$$

The contour plot of the estimated image is provided in Figure 2.4. There, we set the depth of core-mantle boundary (2890 km) as coordinate zero for depth. We can

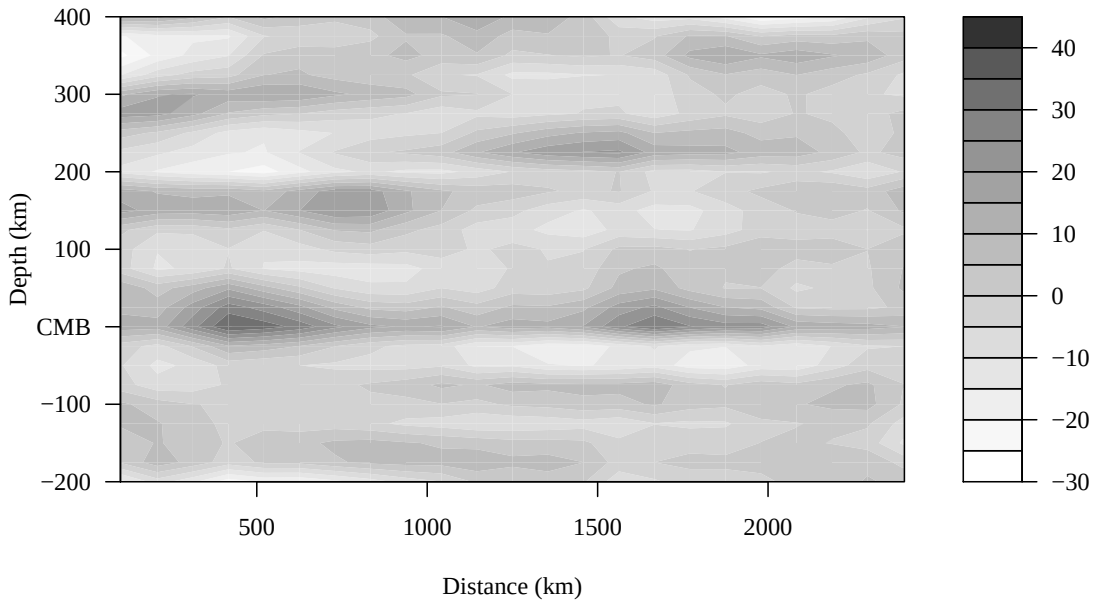

Figure 2.4: The estimated image of core-mantle boundary (CMB) region structure using smoothing spline with adaptive basis sampling.

clearly see a peak at depth zero at all distances, which reveals that the core-mantle boundary is a major boundary. It is interesting that we see two disconnected peaks in the depth around 200 km above the core-mantle boundary: one is below and the other is above. We also calculated 95% Bayesian confidence intervals and found them to indicate that these peaks are significantly nonzero. These structures are likely to be the so-called D'' region, and have also been detected using nonparametric mixed-effect models developed in van der Hilst et al. (2007).

3. REGRESSION WITH RESPONSES FROM EXPONENTIAL FAMILIES

3.1 Introduction

With the rapid development of biotechnologies, second-generation sequencing technologies have become default methods for various genomic and epigenomics analysis, i.e., RNA-seq for gene expression analysis (Mortazavi et al., 2008; Wilhelm et al., 2008; Nagalakshmi et al., 2008), bisulfite sequencing for DNA methylation analysis (Cokus et al., 2008; Lister et al., 2008), and ChIP-seq for genome-wide protein-DNA interaction analysis (Boyer et al., 2005; Johnson et al., 2007; Dixon et al., 2012). Compared to their hybridization-based counterparts, e.g., microarry and ChIP-chip, second generation sequencing technologies offer up to a single-nucleotide resolution signals. In particular, these second-generation sequencing technologies sequence tens of millions of DNA or cDNA fragments in parallel. After mapping the resulting sequences (short reads) to reference genome, researchers get a sequence of read counts. That is, at each nucleotide position, researchers get a count which stands for the number of reads mapped onto that position. By design, these short read counts reflect the quantity of interests. Statistical modeling and inference are indispensable for analyzing the short read counts to facilitate biological discoveries (Li et al., 2010; Ji et al., 2014). Moreover, as the second-generation sequencing technologies become mature and cost-effective, conducting experiments with samples at multiple conditions, and/or of multiple tissue types, and/or at different time points is becoming very common. Since each sequencing sample provides a genome size data, multiple samples give rise to to data of size in tens of millions. The computation of many statistical methods are infeasible to such large sample data. Denote the ith read count by Y_i, which associates with some covariates (features) x_i, where $i = 1, \ldots, n$.

Two typical examples of research works are given below.

3.1.1 Estimating gene expression in RNA-Seq

In these studies, researchers are interested in measuring the quantities of mRNAs molecules, i.e., quantifying gene expressions. Since they are more stable and easily degraded, mRNA molecules are shattered and converted into more stable cDNAs fragments that are short enough suitable for sequencing. The sequenced short fragments are called short reads, which are then aligned to the reference genome to get short-read counts. Finally, gene expressions are estimated based on short-read counts. A simple proposal of estimating gene expression is to average the short-read counts across all nucleotides (within exons) in each gene (normalized by total read counts in the sample) resulting in so-called RPKM (reads per kilobase exon per million mapped reads) (Cloonan et al., 2008). In this approach, the short-read counts at all nucleotides in a gene are assumed to be a *iid* sample of a population. However, significant sequencing bias of short-read counts has been observed (Dohm et al., 2008). In particular, short-read counts at a nucleotide position tend to correlate with GC content in the neighborhood of that nucleotide position. Thus appropriate modeling the variation of short-read counts within each single gene and the variations among genes across the whole genome is crucial to calculating the gene expressions accurately.

Example 1. *Profiling time course gene expression in RNA-Seq.* In these studies, gene expressions over a number of time points in a certain biological process are quantified using RNA-seq. After mapping, read counts at each nucleotide position of the whole genome are obtained at each time point. Thus appropriate modeling the variation of short-read counts within each single gene over time while taking into account the GC bias inherited in the RNA-seq technology is crucial to profiling

the gene expressions over the whole time period accurately. In these studies, the response Y_i is the short-read count of the ith nucleotide in a gene. Besides time t, we also have multivariate factor covariate $x_{it} = (x_{\langle i1t\rangle}, \cdots, x_{\langle iKt\rangle})$, where $x_{\langle ikt\rangle}$ is the GC content in the surrounding k neighborhoods of the ith nucleotide in the gene for $k = 1, \ldots, K$. □

3.1.2 Genome-wide methylation analysis using bisulfite sequencing

DNA methylation is an important epigenetic mechanism that regulates gene expression, cell differentiation and development. It adds a methyl group to a cytosine in CpG dinucleotide (CpG dinucleotide means that a cytosine (C) nucleotide occurs next to a guanine (G) nucleotide. The CpG notation is used to distinguish it from the CG base-pairing in DNA double helix). A current technique for measuring DNA methylation levels is bisulfite sequencing. In this technique, DNA is teated with sodium bisulfate, which converts cytosine (C) residues to uracil (U), but leaves methylated cytosine residue unaffected due to the protection of the methyl group. Hence, bisulphite treatment enables changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single nucleotide resolution information about the methylation status of the DNA sequence (Ji et al., 2014). After sequencing and mapping, the number of short reads mapped onto each CpG site is counted. Thus, bisulfite sequencing data consist of the total number of short reads and methylated reads at each CpG site. Such data allow researchers to estimate methylation proportions at a single-nucleotide resolution.

Example 2. *Identifying differentially methylated regions using bisulfite sequencing.* In these studies , the methylation levels are measured at two conditions using bisulfite sequencing. The goal is to compare the DNA methylation levels and identify the differentially methylated regions (DMRs). After bisulfite sequencing and mapping,

the number of short reads mapped onto each CpG site is counted. The total number of the mapped reads at the ith position is denoted as N_i, and that of methylated reads is denoted as Y_i. To identify the differentially methylated regions, we have bivariate covariate $x_i = (x_{\langle i1\rangle}, x_{\langle i2\rangle})$ where $x_{\langle i1\rangle}$ is the genomic location and $x_{\langle i2\rangle}$ is condition indicator. □

3.1.3 Exponential family smoothing spline ANOVA models

To provide a rich family of distributions in modeling these data, we assume the conditional distribution of Y_i given some covariate x_i has a density in the exponential family with the form

$$f(Y_i|x_i) = \exp\{(Y_i\eta(x_i) - b(\eta(x_i)))/a(\phi) + c(Y_i, \phi)\}, \tag{3.1}$$

where $i = 1, \cdots, n$, $a > 0$, b and c are known functions, $\eta(x)$ is the regression function to be estimated, and ϕ is the dispersion parameter, which is assumed to be a constant, either known or considered as a nuisance parameter. Exponential family includes binomial, Poisson, negative binomial, and log normal distributions in a unified framework and is broad enough to cover all practical applications in second-generation sequencing data.

The short-read counts and their derived data from second-generation sequencing techniques are often in drastically different magnitudes at different genomic positions. Versatile nonparametric modeling of $\eta(x_i)$ in (3.1) provides satisfactory goodness-of-fit (Zheng et al., 2011; Jaffe et al., 2012). Smoothing splines have been primarily employed as simple smoothing tool to remove noise along genomic positions for single sequencing sample (Kuan et al., 2011; Hansen et al., 2012). When one has sequencing data from two treatment groups as in Example 2, smoothing splines may be applied to short read counts of each individual sample separately to get smoothed profiles

along genome positions. Additional models and methods are then applied to the smoothed sequencing profile to extract signal of interest. In principal, such goal can be achieved through an integrated model and inference strategy via analysis of variance through smoothing spline ANOVA (Gu, 2013).

However, the wide application of smoothing spline ANOVA models in modeling second-generation sequencing data has been hindered due to its expensive computational cost, which is $O(n^3)$ where n is sample size. Since the sample size of the second-generation sequencing data is in tens of millions, the computation cost of smoothing spline ANOVA estimates for multivariate x in such super-large samples is prohibitively expensive. Numerous solutions have been proposed in the literature to address the computational issue. For example, hybrid adaptive splines (Luo and Wahba, 1997) integrate a stepwise approach to select nodes and then use the reduced set of nodes to approximate full basis smoothing spline ANOVA models. However, the stepwise nodes selection *per se* is computational expensive. The regression splines (Ruppert et al., 2003) take the advantage of closed form solution of smoothing splines and use a small number of nodes to reduce computational cost. The downside of the method is that nodes placement, in general, needs manually conducted, which is infeasible for genome scale second-generation sequencing data. A recent development along this line of thinking is to use randomly allocated nodes (Gu and Kim, 2002). Albeit it is much simple to implement and faster to compute, the method tends to provide over-smoothed estimates when applied to second-generation sequencing data, and consequently, fails to detect subtle signals.

To surmount these challenges, we develop an adaptive basis selection method for approximating smoothing spline ANOVA models in the exponential family to model second-generation sequencing data. In the proposed method, we evaluate smoothing spline ANOVA models in a lower dimensional effective model space. We construct

the effective model space through an adaptive sampling method via slicing the range of the read counts or the derived data. The sampling strategy and lower effective model space give rise to a more scalable computation for approximating smoothing spline ANOVA models to super-large data, whereas slicing the response provides a representative set of basis functions corresponding different magnitudes of response. The proposed method distinguishes itself from the uniform subsampling approach in selecting smoothing spline basis functions on the direction of the predictors, e.g., Gu and Kim (2002). As evident in our simulation and real data analysis studies, smoothing spline ANOVA models approximation via adaptive basis selection provide very accurate estimates. Our symptomatic theory is nonstandard because of the response-dependent sampling scheme. Our asymptotic functional eigenvalue analysis shows the effective model space is rich enough to retain the essential information of true regression functions and smoothing spline ANOVA models via adaptive basis selection converge at the same convergence rate of regular smoothing spline ANOVA models. Moreover, our theory provides a practical guidelines for choosing the dimension of the effective model space.

The remainder of the chapter is organized as follows. In Section 3.2, we develop the smoothing spline ANOVA via adaptive basis selection method. The asymptotic analysis is presented in Section 3.3. Simulation and real data analysis follow in Sections 3.4 and 3.5. A few remarks in Section 3.6 conclude the chapter. Proofs of the theorems are collected in Chapter 4.

3.2 Efficient computation of smoothing spline ANOVA models via adaptive basis selection

In this section, we first review the penalized likelihood method for fitting smoothing spline ANOVA models and investigate the computation complexity, then develop

the adaptive basis selection method to efficiently approximate the estimator in a low dimensional function space.

3.2.1 Penalized likelihood for fitting smoothing spline ANOVA models

We estimate η by minimizing the penalized likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2}J(\eta), \tag{3.2}$$

where the first term is derived from the negative log likelihood, and $J(\eta) = J(\eta, \eta)$ is a quadratic functional penalizing the roughness of η. When η is a univariate function, a typical choice of J is $J(\eta) = \int(\eta'')^2$. Other examples of J are given at the end of this subsection. The smoothing parameter λ then controls the trade-off between the goodness-of-fit and smoothness of η.

According to (3.1) we can make distribution assumptions in the afore-mentioned examples.

Example 1 (continued) *Profiling time course gene expressions in RNA-Seq.* We assume the short-read count Y given the covariate x is Poisson distributed, i.e., $Y|x \sim$ Poisson($\lambda(x)$) with density $\lambda(x)^Y e^{-\lambda(x)}/Y!$. This is a special case of exponential family density (3.1) with $\eta(x) = \log\lambda(x)$, $a(\phi) = 1$, $b(\eta) = e^\eta$ and $c(Y, \phi) = -\log Y!$. The Poisson intensity $\lambda(x)$ is not to be confused with the smoothing parameter λ appearing in (3.2). Besides Poisson distribution, Sun et al. (2015) models the read count Y by a negative binomial distribution to account for excessive variation in read counts. In particular, $Y|x \sim$ NBinomial$(r, p(x))$ with density $\binom{Y+r-1}{Y}p(x)^Y(1-p(x))^r$ such that $\eta(x) = \log p(x)$, $a(\phi) = 1$, $b(\eta) = -r\log(1-e^\eta)$ and $c(Y, \phi) = \log\binom{Y+r-1}{Y}$ as in (3.1). □

Example 2 (continued) *Identifying differentially methylated regions.* We assume

the number of methylated reads Y given covariate x at position is binomial distributed, i.e., $Y|x \sim \text{Binomial}(N, p(x))$ with density $\binom{N}{Y}p(x)^Y(1-p(x))^{N-Y}$. Compared with (3.1), $\eta(x) = \log\{p(x)/(1-p(x))\}$, $a(\phi) = 1$, $b(\eta) = N\log(1+e^\eta)$ and $c(Y,\phi) = \log\binom{N}{Y}$. $\square$

Usually the quadratic functional $J(\eta)$ is a square semi-norm and the standard formulation of smoothing splines restricts minimizing (3.2) in a reproducing kernel Hilbert space (RKHS) $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. A Hilbert space has a metric and a geometry that facilitate analysis and computation. To prevent interpolation, the null space of J, $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$, is assumed to be a finite dimensional linear subspace of $\mathcal{H}$ with basis $\{\phi_i\colon\ i = 1, \cdots, m\}$. Denote the orthogonal decomposition of $\mathcal{H}$ by $\mathcal{N}_J \oplus \mathcal{H}_J$ where $\mathcal{H}_J$ is still a reproducing kernel Hilbert space. Let $R_J(x,y)$ be the reproducing kernel of $\mathcal{H}_J$. The representer theorem (Wahba, 1990) shows that the minimizer of (3.2) in the RKHS $\mathcal{H}$ have a simple form

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R_J(x_i, x), \tag{3.3}$$

where coefficients d_ν and c_i are to be estimated from data.

For multivariate x, the smoothing spline analysis of variance (ANOVA) decomposition of a multivariate function η is

$$\eta(x) = \eta_0 + \sum_{j=1}^{d}\eta_j(x_{\langle j\rangle}) + \sum_{j=1}^{d}\sum_{k=j+1}^{d}\eta_{jk}(x_{\langle j\rangle}, x_{\langle k\rangle}) + \ldots + \eta_{1,\ldots,d}(x_{\langle 1\rangle}, \ldots, x_{\langle d\rangle}) \tag{3.4}$$

where the η_0 is a constant, the η_j's are the main effects, the η_{jk}'s are the two-way interactions, etc. The identifiability of the terms in (3.4) is ensured by side conditions through averaging operators (Wahba, 1990; Gu, 2013). To use (3.4) for estimating η in (3.2) , we consider $\eta_j \in \mathcal{H}_{\langle j\rangle}$, where $\mathcal{H}_{\langle j\rangle}$ is an RKHS with tensor sum

decomposition $\mathcal{H}_{\langle j\rangle} = \mathcal{H}_{0\langle j\rangle} \oplus \mathcal{H}_{1\langle j\rangle}$, where $\mathcal{H}_{0\langle j\rangle}$ is the finite-dimensional "parametric" subspace consisting of parametric functions, and $\mathcal{H}_{1\langle j\rangle}$ is the "nonparametric" subspace consisting of smooth functions. The induced tensor product space is

$$\mathcal{H} = \otimes_{j=1}^{d} \mathcal{H}_{\langle j\rangle} = \oplus_{\mathcal{S}}[(\otimes_{j\in\mathcal{S}} \mathcal{H}_{1\langle j\rangle}) \otimes (\otimes_{j\notin\mathcal{S}} \mathcal{H}_{0\langle j\rangle})] = \oplus_{\mathcal{S}} \mathcal{H}_{\mathcal{S}},$$

where the summation runs over all subsets $\mathcal{S} \subseteq \{1, \ldots, d\}$. The corresponding penalty function $J(\eta) = \sum_{\mathcal{S}} \theta_{\mathcal{S}}^{-1} J_{\mathcal{S}}(\eta_{\mathcal{S}})$ with $\eta_{\mathcal{S}} \in \mathcal{H}_{\mathcal{S}}$, $\theta_{\mathcal{S}} > 0$ are extra smoothing parameters, and $J_{\mathcal{S}}$ is the square norm in $\mathcal{H}_{\mathcal{S}}$. The subspaces $\mathcal{H}_{\mathcal{S}}$ form two large subspaces: $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$, which is the null space of $J(\eta)$, and $\mathcal{H} \ominus \mathcal{N}_J$ with the reproducing kernel $R_J = \sum_{\mathcal{S}} \theta_{\mathcal{S}} R_{\mathcal{S}}$ where $R_{\mathcal{S}}$ is the reproducing kernel in $\mathcal{H}_{\mathcal{S}}$. The smoothing spline estimator in such reproducing kernel Hilbert space is called a tensor product smoothing spline.

Example 1 (continued) *Profiling time course gene expressions in RNA-Seq.* In the functional ANOVA decomposition of $\eta(x)$, $\exp\{\eta_0\}$ denotes time course gene expression level, all other main effects and interactions are sequencing bias need to be removed. □

Example 2 (continued) *Identifying differentially methylated regions.* Applying function ANOVA to η, we can identify significantly different methyaltion profiles over two conditions. □

In Example 2, we encounter the case where covariates are of mixed types. Consider a bivariate function $\eta(x, \tau)$, where $x \in [0, 1]$ and $\tau \in \{1, \ldots, t\}$. A valid decomposition is $\eta(x, \tau) = \eta_\emptyset + \eta_1(x) + \eta_2(\tau) + \eta_{1,2}(x, \tau)$, where $\eta_\emptyset$ is a constant, $\eta_1(x)$ is a function of x satisfying $\eta_1(0) = 0$, $\eta_2(\tau)$ is a function of τ satisfying $\sum_{\tau=1}^{t} \eta_2(\tau) = 0$, and $\eta_{1,2}(x, \tau)$ satisfies $\eta_{1,2}(0, \tau) = 0$, $\forall\tau$, and $\sum_{\tau=1}^{t} \eta_{1,2}(x, \tau) = 0$, $\forall x$. Regarding the

quadratic functional J, one may use

$$J(\eta) = \theta_1^{-1} \int_0^1 (d^2\eta_1/dx^2)^2 dx + \theta_{1,2}^{-1} \int_0^1 \sum_{\tau=1}^{t} (d^2\eta_{1,2}/dx^2)^2 dx.$$

The null space $\mathcal{N}_J$ has dimension $2t$ with basis given by

$$\{1, x, I_{[\tau=j]} - 1/t, (I_{[\tau=j]} - 1/t)x, j = 1, \ldots, t-1\}.$$

Moreover, the reproducing kernel of $\mathcal{H}_J$ is

$$R_J(x_1, \tau_1; x_2, \tau_2) = \theta_1 \int_0^a (x_1 - u)_+ (x_2 - u)_+ du + \theta_{1,2}(I_{[\tau_1=\tau_2]} - 1/t) \int_0^a (x_1 - u)_+ (x_2 - u)_+ du.$$

General discussion can be found in Chapter 2.4 of Gu (2013).

By standard exponential family theory, $\mathrm{E}[Y|x] = b'(\eta(x)) = \mu(x)$ and $\mathrm{var}[Y|x] = b''(\eta(x))a(\phi) = \nu(x)a(\phi)$. When the likelihood function in model (3.2) has a unique minimizer in $\mathcal{N}_J$, the minimizer $\hat{\eta}$ of (3.2) uniquely exists. Fixing the smoothing parameter λ (and ones hidden in $J(\eta)$, if present), (3.2) may be minimized through the Newton iteration. Write $l(\eta(x_i); Y_i) = -Y_i\eta(x_i) + b(\eta(x_i))$, $u(\eta(x_i); Y_i) = -Y_i + b'(\eta(x_i))$, and $w(\eta(x_i); Y_i) = b''(\eta(x_i)) = \nu(x_i)$. The quadratic approximation of $l(\eta(x_i); Y_i)$ at the current estimate $\tilde{\eta}(x_i)$ is seen to be

$$l(\eta(x_i; Y_i) \approx l(\tilde{\eta}(x_i); Y_i) + \tilde{u}_i(\eta(x_i) - \tilde{\eta}(x_i)) + \tilde{w}_i(\eta(x_i) - \tilde{\eta}(x_i))^2/2 = \tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2/2 + C_i,$$

where $\tilde{u}_i = u(\tilde{\eta}(x_i); Y_i)$, $\tilde{w}_i = w(\tilde{\eta}(x_i); Y_i)$, $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$ and C_i is independent of $\eta(x_i)$. The Newton iteration can thus be performed to penalized weighted least squares,

$$\sum_{i=1}^{n} \tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + n\lambda J(\eta). \tag{3.5}$$

Although fast algorithms (Reinsch, 1967) are available when x is univariate, the computation of (3.5) for multivariate x is at least in the order of $O(n^3)$, see Chapter 3.4 of Gu (2013). The high computational cost of smoothing splines render its inapplicability in modeling second-generation sequencing data. In our examples, sample sizes are 48,660 and 23,361.

3.2.2 Adaptive basis selection

To alleviate computational cost of smoothing splines, one may restrict the minimizer of (3.2), equivalently (3.5), in a reduced subspace of $\mathcal{H}$. Such subspace is called an effective model space with two distinguishing features. First, the computational cost in constructing the effective model space is very inexpensive; second, the effective model space retains the essential information of the true function η. Gu and Kim (2002) and Kim and Gu (2004) developed a simple random sampling approach to select a subset of full basis functions and construct an effective model space. The resulted estimator shares the same asymptotic convergence rates with the estimator constructed with full basis functions. Following this idea, we propose an adaptive basis selection algorithm to construct the effective model space. The intuition is to first produce a crude estimate of the conditional density $f(x|y)$ using a simple slicing technique and then guarantee the most influential basis functions included in the effective model space. In particular, when the underlying function varies significantly in magnitude, the estimates based on such adaptive sampling approach will outperform the the estimates based on uniform random sampling.

Adaptive basis selection algorithm

(1) Divide the range of the responses $\{Y_i\}_{i=1}^n$ *into a number of disjoint intervals, say* K *intervals, which are denoted by* S_1, S_2, ..., S_K.

(2) *For $k = 1, \ldots, K$, take a random sample $x_1^{*(k)}, \ldots, x_{n_k}^{*(k)}$ of size n_k without replacement, from original sample x_i with probability $|S_k|^{-1} I_{y_i \in S_k}$, where $|S_k|$ is the number of observations in S_k. We denote the combined sample as $x_1^*, \ldots, x_{n^*}^*$ with sample size n^*.*

(3) *Finally, minimizing criterion (3.2) over*

$$\mathcal{H}_E = \mathcal{N}_J \oplus span\{R_J(x_j^*, \cdot), j = 1, \ldots, n^*\}$$

where $\mathcal{H}_E$ is referred to as the effective model space. The minimizer then has the expression

$$\hat{\eta}_A(x) = \sum_{i=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{n^*} c_j R_J(x_j^*, x) \tag{3.6}$$

where $\hat{\eta}_A(x)$ is a smoothing spline ANOVA estimate through adaptive basis selection.

When dividing the range of response variable, we need take the specific exponential family into account. In Example 1, responses follow a Poisson distribution and we can apply step (1) directly. However, in Example 2 where Y_i follows a binomial distribution, we instead propose to divide the range of ratio Y_i/N_i to avoid possible heterogeneity in count data.

With the adaptively selected basis, one can reformulate the minimization of the penalized weighted least squares functional (3.5). Substituting (3.6) into (3.5), the numerical problem becomes minimizing

$$(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c})^T \tilde{W} (\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c} \tag{3.7}$$

with respect to $\mathbf{d}$, $\mathbf{c}$, where $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, S is $n \times m$ with the (i, ν)th entry

$\phi_\nu(x_i)$, R is $n \times n^*$ with the (i,j)th entry $R_J(x_i, x_j^*)$, Q is $n^* \times n^*$ with the (j,k)th entry $R_J(x_j^*, x_k^*)$, and $\tilde{W} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$. The solution of (3.7) satisfies the normal equation

$$\begin{pmatrix} S_w^T S_w & S_w^T R_w \\ R_w^T R_w & R_w^T R_w + (n\lambda)Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S_w^T \tilde{\mathbf{Y}}_w \\ R_w^T \tilde{\mathbf{Y}}_w \end{pmatrix}, \tag{3.8}$$

where $S_w = \tilde{W}^{1/2} S$, $R_w = \tilde{W}^{1/2} R$, and $\tilde{\mathbf{Y}}_w = \tilde{W}^{1/2} \tilde{\mathbf{Y}}$. The normal equation of (3.8) can be solved by the pivoted Cholesky decomposition followed by backward and forward substitutions (Kim and Gu, 2004).

On the convergence of Newton iteration, the "fitted values" $\hat{\mathbf{Y}}_w = S_w \mathbf{d} + R_w \mathbf{c}$ of (3.5) can be written as $\hat{\mathbf{Y}}_w = A_w(\lambda) \tilde{\mathbf{Y}}_w$, where the smoothing matrix

$$A_w(\lambda) = (S_w, R_w) \begin{pmatrix} S_w^T S_w & S_w^T R_w \\ R_w^T R_w & R_w^T R_w + (n\lambda)Q \end{pmatrix}^+ \begin{pmatrix} S_w^T \\ R_w^T \end{pmatrix}.$$

and $\mathbf{C}^+$ denotes the Moore-Penrose inverse of $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}$, $\mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+$, $(\mathbf{C}\mathbf{C}^+)^T = \mathbf{C}\mathbf{C}^+$ and $(\mathbf{C}^+\mathbf{C})^T = \mathbf{C}^+\mathbf{C}$.

A data-driven approach for the selection of the tuning parameter λ (including θ) is to choose λ which minimizes the generalized approximate cross-validation score (Gu and Xiang, 2001),

$$GACV(\lambda) = -\frac{1}{n} \sum_{i=1}^{n} \{Y_i \hat{\eta}_A(x_i) - b(\hat{\eta}_A(x_i))\} + \frac{\mathrm{tr}(A_w \tilde{W}^{-1})}{n - \mathrm{tr} A_w} \frac{1}{n} \sum_{i=1}^{n} Y_i (Y_i - \hat{\mu}(x_i)). \tag{3.9}$$

One may employ standard nonlinear optimization algorithms to minimize the generalized approximate cross-validation score. In particular, we use the modified Newton algorithm developed by Dennis and Schnabel (1996) to find the minimizer. $\hat{\eta}_A$ and

$\hat{\mu}$ are evaluated at the minimizer of (3.2) with fixed tuning parameters, and A_w and $\tilde{W}$ are evaluated on the convergence of Newton iteration.

3.3 Asymptotic analysis

We now develop an asymptotic analysis to guide the construction of the effective model space and establish the convergence rate of smoothing spline with adaptive basis selection. Since our basis selection algorithm involves the response variable, the standard argument for the asymptotic analysis of smoothing splines does not apply. We refer to Chapter 4 for some theoretical properties of adaptive basis sampling which shed light on how the algorithm works and facilitate our asymptotic analysis

3.3.1 Regularity conditions

Recall that $l(\eta(x); y) = -y\eta(x) + b(\eta(x))$ and $u(\eta; y) = dl/d\eta$, $w(\eta; y) = d^2l/d\eta^2$ and assume that

$$\mathrm{E}\{u(\eta_0(X); Y)|X\} = 0, \quad \mathrm{E}\{u^2(\eta_0(X); Y)|X\} = \sigma^2 \,\mathrm{E}\{w(\eta_0(X); Y)|X\}.$$

Write $v_\eta(x) = \mathrm{E}\{w(\eta(x); Y)|X = x\}$. Let $f_X(\cdot)$ be the marginal density of the predictor variable X and define

$$V(g) = \int_{\mathcal{X}} g^2(x) v_{\eta_0}(x) f_X(x)\, dx.$$

Condition 3.1. *V is completely continuous with respect to J.*

This condition ensures that there exists a sequence of eigenfunctions $\phi_\nu \in \mathcal{H}$ and the associated nonnegative increasing sequence of eigenvalues ρ_ν such that functionals V and J are simultaneously diagonalized. That is, $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$ where $\delta_{\nu\mu}$ is the Kronecker delta. Furthermore, any function f satisfying

$J(f) < \infty$ can be expressed as a Fourier series expansion $f = \sum_\nu f_\nu \phi_\nu$, where $f_\nu = V(f, \phi_\nu)$.

Condition 3.2. *For some $r > 1$ and $\beta > 0$, we have $\rho_\nu > \beta\nu^r$ for sufficiently large ν.*

The growth rate of the eigenvalues ρ_ν of J with respect to V, essentially dictates how fast λ should approach to zero. Such polynomial rate is satisfied in various models, including polynomial splines, thin-plate splines and spherical splines. See Chapter 9.1 of Gu (2013).

Condition 3.3. *For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$,*

$$c_1\, w(\eta_0(x); y) \leqslant w(\eta(x); y) \leqslant c_2\, w(\eta_0(x); y)$$

holds uniformly for some $0 < c_1 < c_2 < \infty$, $\forall x \in \mathcal{X}$, $\forall y$.

Roughly speaking, Condition 3.3 concerns the equivalence of the information within B_0.

Condition 3.4. *There is a constant $c_3 < \infty$ such that $var\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X); Y)\} \leqslant c_3$ for all ν, μ.*

Recall that ϕ_ν's forms an orthonormal system relative to $V(\cdot, \cdot)$ such that

$$\mathrm{E}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X); Y)\} = V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu},$$

and thus $\mathrm{E}\{\phi_\nu^2(X)\phi_\mu^2(X)w^2(\eta_0(X); Y)\} \leqslant c_3 + 1$. Condition 3.4 basically requires the fourth moments of $\phi_\nu(X)$ is uniformly bounded.

3.3.2 Rate of convergence

For completeness, we first state a standard result for the convergence rate of smoothing splines with full basis, $\hat{\eta}$ as in (3.3). The following result is Theorem 9.17 in Gu (2013).

Theorem 3.3.1. *Assume that* $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ *for some* $p \in [1, 2]$. *Under Condition 3.1-3.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, *we have*

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

To understand the behavior of $\hat{\eta}_A$, the smoothing spline estimator computed using the adaptive basis selection algorithm, we first show two important properties of the effective model space $\mathcal{H}_E$.

Lemma 3.3.1. *For any function outside the effective model space, its evaluations at selected samples* $\{x_j^*\}_{j=1}^{n^*}$ *are all zeros, i.e. for* $h \in \mathcal{H} \ominus \mathcal{H}_E$,

$$h(x_j^*) = 0, \qquad j = 1, \ldots, n^*.$$

Lemma 3.3.2. *Under Condition 3.1, 3.2, and 3.4, as* $\lambda \to 0$ *and* $n^*\lambda^{2/r} \to \infty$, *if function* h *is not in the effective model space, i.e.,* $h \in \mathcal{H} \ominus \mathcal{H}_E$, *we have*

$$V(h) = o_p\{\lambda J(h)\}.$$

We now present our main result on the convergence rate of $\hat{\eta}_A$.

Theorem 3.3.2. *Assuming that* $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ *for some* $p \in [1, 2]$. *Under*

Condition 3.1-3.4, as $\lambda \to 0$ *and* $n^*\lambda^{2/r} \to \infty$*, we have*

$$(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

In particular, when $\lambda \asymp n^{-r/(pr+1)}$*, the estimator* $\hat{\eta}_A$ *achieves the optimal convergence rate*

$$(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p(n^{-pr/(pr+1)}).$$

This theorem states that, under regularity conditions, the convergence rate of the smoothing spline estimator using an adaptively selected basis is the same as that of the smoothing spline estimator using the full basis indicated by the representer theorem.

3.3.3 The dimension of the effective model space

Utilizing the asymptotic analysis results in the last subsection, we now determine the dimension of the effective model space $\mathcal{H}_E$. On one hand, with $\lambda \asymp n^{-r/(pr+1)}$, Lemma 3.3.2 and Theorem 3.3.2 both require $n^*\lambda^{2/r} \to \infty$, which implies $n^* \asymp n^{2/(pr+1)+\delta}$, where δ is an arbitrary small positive number. On the other hand, constant p depends on the smoothness of η: for roughest η satisfying $J(\eta) < \infty$, we have $p = 1$, whereas for the smoothest η, we have $p = 2$.

Take univariate cubic smoothing spline as an example: $J(\eta) = \int_0^1 (\eta'')^2$ with $r = 4$ and $\lambda \asymp n^{-4/(4p+1)}$. The proper dimension of the effective model space is $n^* = n^{2/(4p+1)} + \delta$, which is in the range of $O(n^{2/9+\delta})$ and $O(n^{2/5+\delta})$ for p in $[1, 2]$. For linear smoothing spline, the range is $(O(n^{2/5+\delta}), O(n^{2/3+\delta}))$ In our simulation and example, we take dimension of the effective model space n^* to be between $4n^{2/9}$ and $20n^{2/9}$ for cubic smoothing spline with selected basis, between $4n^{2/5}$ and $20n^{2/5}$ for linear smoothing spline with selected basis.

3.4 Simulation study

We approximated smoothing spline ANOVA estimate via adaptive basis sampling and that with uniform basis sampling (Kim and Gu, 2004) to three multivariate test functions. Exponential family distributions considered include negative binomial, Poisson and binomial . In generating predictors x, a random design was adopted: $n = 1600$ points were uniformly generated from the domains. Responses were correspondingly generated under each distribution assumption. The number of slices was suggested by Scott's method (Scott, 1992) and based on our asymptotic results, the dimension of the effective model space was set to be $10n^{2/9}$, which meant $n^* = 52$ basis functions were sampled for both sampling methods for approximating smoothing spline ANOVA models.

We first took the bivariate blocks function with negative binomial distribution as an example. The bivariate blocks function is a direct generalization of the univariate blocks function (Donoho and Johnstone, 1994) to two dimensional. Let blocks$(\cdot)$ be the univariate blocks function, then the bivariate blocks function is $\text{blocks}_2(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = \text{blocks}(x_{\langle 1\rangle})$, For the negative binomial distribution with parameters (α, p), we set the success probability $p = (\text{blocks}_2 + 2.5)/8$ and the target for number of successful trials $\alpha = 3$.

The next two examples were constructed from the joint probability density of a d-dimensional nonparanormal distribution (Liu et al., 2009), which is given by

$$p_\alpha^d(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} f(x)^\top \Sigma^{-1} f(x)\right\} \prod_{j=1}^{d} |f_j'(x_j)|, \tag{3.10}$$

where Σ is a $d \times d$ matrix with diagonal entries to be 1, super and sub diagonal entries to be 0.5 and other entries to be 0, and the jth component of $f(x)$ takes the

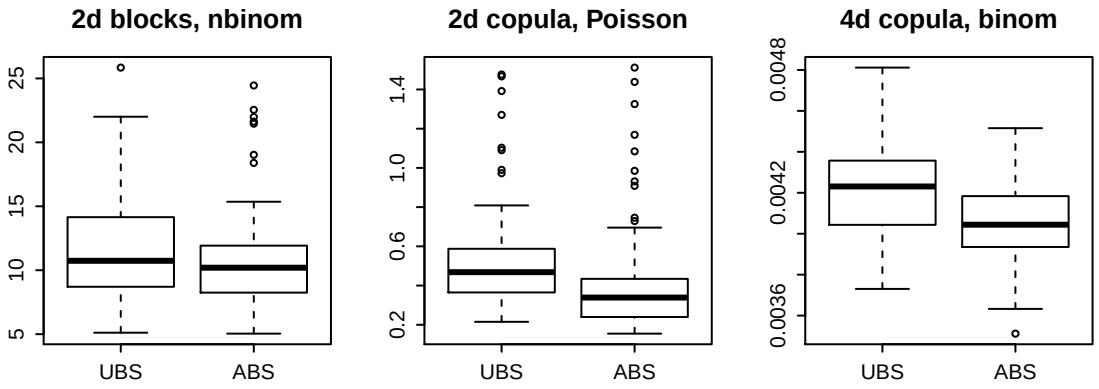

Figure 3.1: Boxplots of MSE for multivariate simulation studies. Left: bivariate blocks function with negative binomial distribution; middle: bivariate copula density function with Poisson distribution; right: four dimensional copula density function with binomial distribution. UBS and ABS stand for smoothing spline ANOVA models estimator under uniform and adaptive basis sampling strategies.

form $f_j(x) = \alpha_j \operatorname{sign}(x) |x|^{\alpha_j}$ and α_j's are shape parameters.

The second example was a bivariate copula density function with Poisson distribution. The bivariate copula density was obtained by setting $d = 2$ and $\alpha = (2, 3)^\top$ in (3.10). For Poisson distribution, the mean parameter $\lambda = 1 + 2(2\pi)^{p/2}|\Sigma|^{1/2}\, p_\alpha^d$. Our third example was a higher dimensional example, a four dimensional copula density function with binomial distribution. Let $d = 4$ and $\alpha = (0.1, 0.1, 0.1, 0.1)^\top$ in (3.10). For binomial distribution with parameters (m, p), the number of trials $m = 50$ and the success probability $p = \exp(p_\alpha^d)/\{1 + \exp(p_\alpha^d)\}$.

To evaluate the performance of each approximation method, we repeated the experiment for 100 times under each simulation set-up and calculated the mean squared error (MSE) for the estimate. For binomial and negative binomial distribu-

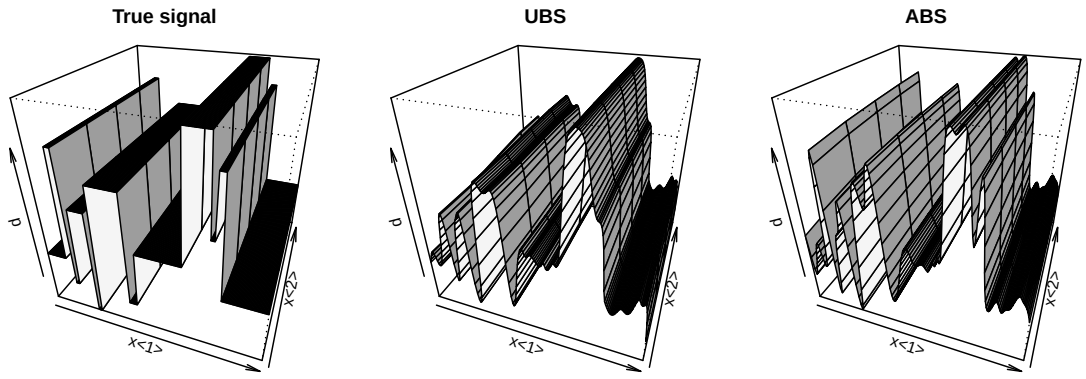

Figure 3.2: Bivariate blocks function with negative binomial distribution. Perspective plots of true probability, fitted values by smoothing splines via uniform basis sampling and adaptive basis sampling.

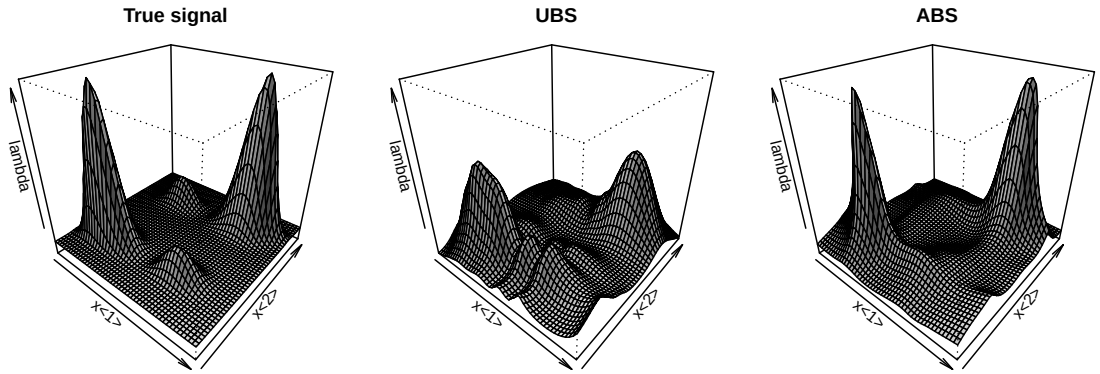

Figure 3.3: Bivariate copula density function with Poisson distribution. Perspective plots of true mean parameter, fitted values by smoothing splines via uniform basis sampling and adaptive basis sampling.

tions MSE $= \sum_{i=1}^{n}\{\hat{p}(x_i) - p(x_i)\}^2$ and for Poisson distribution MSE $= \sum_{i=1}^{n}\{\hat{\lambda}(x_i) - \lambda(x_i)\}^2$. Boxplots of MSEs for three multivariate test functions are displayed in Figure 3.1. It is obvious that the proposed adaptive basis sampling scheme enables smoothing spline ANOVA models to be more accurate and stable. Further calculation shows that, under the three simulation set-ups smoothing splines with adaptive basis sampling outperforms that with uniform basis sampling for 69, 96 and 76 times out of 100 experiments respectively.

Figure 3.2 and 3.3 display the visualization for two 2-dimensional examples for a single run. In Figure 3.2, the probability parameter of the negative binomial distribution is a bivariate blocks function which has many abrupt local jumps on $x_{\langle 1\rangle}$ direction. The proposed method successfully recovers those fine scale information while uniform basis sampling fails. In Figure 3.3, the mean parameter of the Poisson distribution behaves relatively smooth. There are four peaks across the domain: two are significantly higher than the other two. Smoothing splines with adaptively sampled basis apparently provides a better estimate: the two big peaks recovered are closer to the truth.

3.5 Real examples

In this section, we analyze two sequencing data sets from Examples 1 and 2 respectively.

3.5.1 Modeling the time course gene expression profiles using RNA-Seq

Drosophila melanogaster (fruit fly) shares a substantial genetic content with humans and has been used as a translational model for human development. To study Drosophila melanogaster development, Graveley et al. (2011) conducted time course RNA-seq experiments. In these experiments, the authors collected 12 embryonic RNA samples at two-hour intervals for 24 hours in the stage of early embryos. The

samples were then sequenced using an Illumina Genome Analyzer IIx platform.

To enhance our understanding of gene expression dynamics, we are interested in estimating time courses gene expressions at the early embryos stage. To estimate time course gene expressions accurately, we need to take into account the sequencing bias, in particular the GC bias. To do this, we attempt a nonparametric model to model time course gene expression profiles while accounting for the GC bias. Since the read in Graveley et al. (2011) is 76 base-pair long, we count the GC content in each read length interval. We denote short-read counts at the jth nucleotide of the ith gene at time point t by Y_{ijt}, the number of GC counts in the neighborhood of 1 to 76 nucleotides away from the jth nucleotide by $x_{\langle 1ij\rangle}$, that in the neighborhood of 77 to 152 nucleotides away from the jth nucleotide by $x_{\langle 2ij\rangle}$, and that in the neighborhood of 153 to 228 nucleotides away from the jth nucleotide by $x_{\langle 3ij\rangle}$. We built a Poisson nonparametric model for the short-read counts,

$$Y_{ijt} \sim \text{Poisson}(\lambda_{ijt}),$$

the mean λ_{ijt} follows the following model,

$$\log(\lambda_{ijt}) = \alpha_i + \eta_0(t) + \eta(x_{\langle 1ij\rangle}, x_{\langle 2ij\rangle}, x_{\langle 3ij\rangle}), \tag{3.11}$$

where baseline gene expression level of the ith gene is $\exp\{\alpha_i\}$, $\eta_0(t)$ is time trend, and $\eta(x_{\langle 1ij\rangle}, x_{\langle 2ij\rangle}, x_{\langle 3ij\rangle})$ is the sequencing bias. We then applied the smoothing spline ANOVA decomposition to $\eta(x_{\langle 1\rangle}, x_{\langle 2\rangle}, x_{\langle 3\rangle})$. In the smoothing spline ANOVA models, we kept all main effects, two-way and three-way interactions of covariates,

i.e.,

$$\eta(x_{\langle 1\rangle}, x_{\langle 2\rangle}, x_{\langle 3\rangle}) = c_0 + \sum_{k=1}^{3} \eta_k(x_{\langle k\rangle}) + \sum_{k=1}^{3} \sum_{l=j+1}^{3} \eta_{kl}(x_{\langle k\rangle}, x_{\langle l\rangle}) + \eta_{123}(x_{\langle 1\rangle}, x_{\langle 2\rangle}, x_{\langle 3\rangle}). \tag{3.12}$$

We then fit the smoothing spline ANOVA models using penalized likelihood (3.2) for all genes. Since most genes have several thousand nucleotides, the total number of observations n is around $50,000$, which renders the fitting of smoothing spline ANOVA models infeasible. Instead, we fitted smoothing spline ANOVA models using the proposed adaptive basis sampling method.

Here, we illustrate the analysis of RNA-seq data using two randomly selected genes. The selected genes are heat shock protein cognate 4 (Hsc70-4) and elongation factor 2b (Ef2b), which are $3,974$ and $4,055$ bp long (only exons are kept) respectively. Using our adaptive basis selection method for fitting smoothing spline ANOVA models, we set dimension of effective modeling space (the number of basis) $n^* = 72$ for both genes. The computing time for running the smoothing spline ANOVA models with adaptive basis selection are 95 and 124 CPU seconds on a 2.90 GHz Intel Xeon computer. To further assess the adequacy of the smoothing splines ANOVA estimates via adaptive basis selection, we computed the quasi-R^2 (Li et al., 2010), which is defined as

$$R^2 = 1 - d/d_0 \tag{3.13}$$

where d is the deviance of the fitted model and d_0 is the deviance of the null model with only constant mean. The quasi-R^2 of the fitted smoothing spline ANOVA model via adaptive sampling method is 0.87 for Hsc70-4, and 0.86 for Ef2b. Figure 3.4 and 3.5 display the estimated counts $\exp \alpha_i + \eta_0(t)$ by removing GC bias η from λ_{ijt} in two genes.

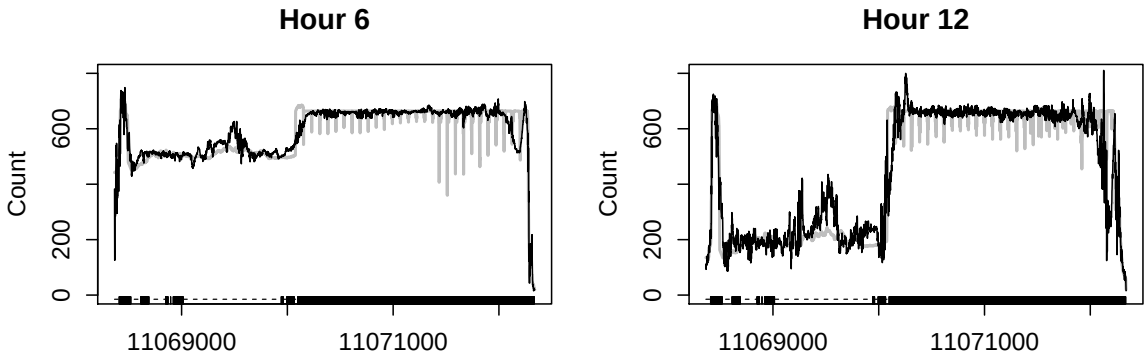

Figure 3.4: Estimated counts after removing GC bias for two time courses of gene Hsc70-4. Observed counts are in gray line and black line is the estimation, the blocks in the bottom are exons.

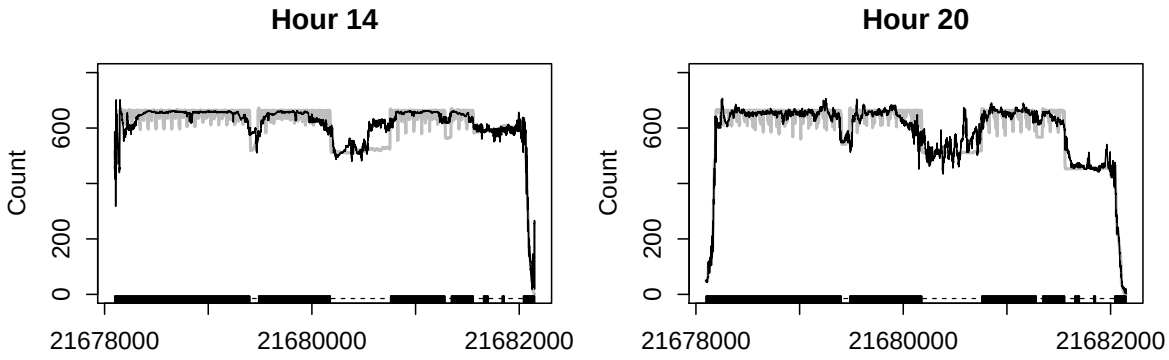

Figure 3.5: Predicted counts after removing GC bias for two time courses of gene Ef2b. Observed counts are in gray line and black line is the estimation, the blocks in the bottom are exons.

Since RNA-seq data provide single nucleotide data, we can further estimate the isoform gene expressions (Jiang and Wong, 2009). According to flybase annotation (www.flybase.org), there are seven known isoforms for Hsc70-4 gene and three for Ef2b. We estimated the isoform expression for both Hsc70-4 and Ef2b. The estimated isoform expressions at Hour 6 and 12 for Hsc70-4 and those for Ef2b at Hour 14 and 20 are listed in Table 3.1 and 3.2

Table 3.1: Raw read counts and fitted counts for all 7 isoforms of gene Hsc70-4 at Hour 6 and 12.

	Hour 6					
	Isoform 1	Isoform 2	Isoform 3	Isoform 4	Isoform 5	Isoform 6
Raw	1522416	1492814	1466414	1468447	1503038	1495502
Fitted	1503707	1474983	1446416	1448412	1482390	1476911
	Isoform 7					
Raw	1506437					
Fitted	1488867					
	Hour 12					
	Isoform 1	Isoform 2	Isoform 3	Isoform 4	Isoform 5	Isoform 6
Raw	1486773	1443093	1435028	1435856	1450492	1445258
Fitted	1441400	1399584	1388391	1389162	1401834	1401303
	Isoform 7					
Raw	1450323					
Fitted	1408558					

Table 3.2: Raw read counts and fitted counts for all 3 isoforms of gene Ef2b at Hour 14 and 20.

	Hour 14			Hour 20		
	Isoform 1	Isoform 2	Isoform 3	Isoform 1	Isoform 2	Isoform 3
Raw	1824904	1809689	1824718	1773156	1766892	1778588
Fitted	1798631	1781373	1796776	1749174	1742474	1754925

3.5.2 Differentially methylated DNA regions in Arabidopsis

DNA methylation is an important epigenetic mechanism that regulates gene expression, cell differentiation and development. The whole genome GC methylation levels of four strains of *Arabidopsis thaliana* were measured using whole genome bisulfite sequencing (Ji et al., 2014). The whole genome of Arabidopsis is around 135 million bp. Two strains were from one generation and the other two strains were taken from a second generation. The total number of GC methylated nucleotides is $23,361$.

Let $Y_{i,s,g}$ be the read counts at genetic position i in strain s of generation g, where $s = 1, 2$ and $g = 1, 2$. We build a binomial nonparametric model for the short-read counts, $Y_{i,s,g} \sim \text{Binomial}(N_{i,s,g}, p(i,s,g))$. The canonical parameter is

$$\log \frac{p(i,s,g)}{1-p(i,s,g)} = \eta(i,g) + b_s,$$

where η is further decomposed through a smoothing spline ANOVA decomposition,

$$\eta(i,g) = \eta_0 + \eta_1(i) + \eta_2(g) + \eta_{12}(i,g),$$

and random effect $b_s \sim N(0, \sigma^2)$ induces the spatial correlation in the methylation for each strain.

Since the genome of *Abrabidopsis* is around 135 million bp, we divided the whole genome into 20 k bp segments and fit the model to each segments using our adaptive basis selection method. In the DNA methylation data, we observe that short-read count $N_{i,s,g}$ varies significantly with position s. Hence, when applying our adaptive basis selection method, we first divided the range of the ratio of methylated read counts to total read counts to disjoint intervals. Thus we selected the basis using

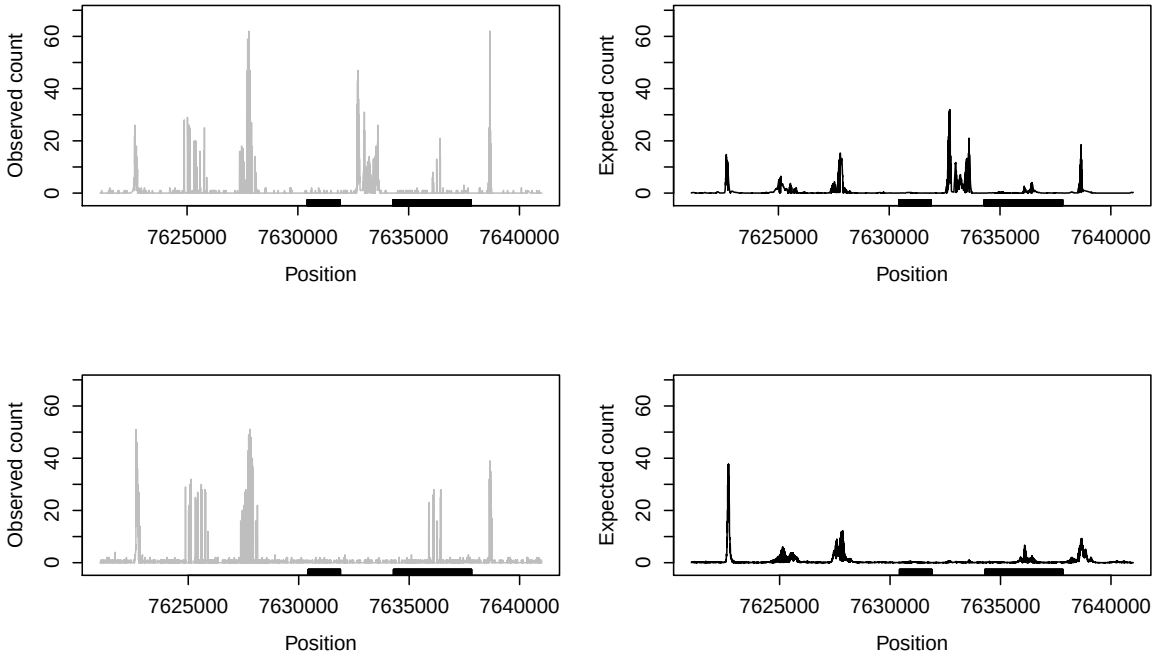

Figure 3.6: Mapped methylated read counts and fitted methylation level for a whole genome bisulfite sequencing data of *Arabidopsis thaliana*. The grey lines at left panels are the mapped methylation read counts for four strains of two generations. The black lines are the fitted methylation levels. The thick bars in x-axises are location of genes AT2G17540 (left) and AT2G17550 (right)

an empirical estimate of success probability $p(i, g)$. The size of the effective model space is controlled as $n^* = 102$ and the CPU running time is about ten minutes in a computer with an Intel Xeon 2.90 GHz processor with 64GB of DDR3 RAM. We then test the significance of $\eta_{12}(i, g)$ and $\eta_2(g)$ using Kullback-Leibler projection (see Section 5.3 (Gu, 2013)) to identify differentially methylated regions.

An identified DMR region is plotted in Figure 3.6. This DMR is in chromosome 2 ranging from 7621000 to 7641000. The Kullback-Leibler ratio projection for position only model is 0.11, for position and generation additive model is 0.08. In other words, the Kullback-Leibler projections suggest that none of the three terms

can be eliminated. In particular, the DMR region is in the intergenic region between gene AT2G17540 and gene AT2G17550 (TON1 RECRUITING MOTIF 26, TRM26). Organization of the cortical cytoskeleton guides the growth and morphogenesis of organisms, e.g., *Arabidopsis*, that depend on cell walls. By positioning wall-building enzymes, the cytoskeleton acts as an interior scaffold to direct construction of the cell's exterior. In plants, environmental and hormonal signals that modulate cell growth cause reorganization of cortical microtubule arrays (Lindeboom et al., 2013). It has been conformed that in *Arabidopsis thaliana*, the TON1 proteins are essential for microtubule organization at the cortex (Drevensek et al., 2012). Thus, the identified DMR region is likely to be concertedly worked with TON1 protein to regulated microtubule organization.

3.6 Discussion

Proper modeling of second-generation sequencing data plays an important role in navigating the biological discovery. In this article, we developed an effective approximation of smoothing spline ANOVA via adaptive basis selection for nonparametric modeling of second-generation sequencing data. Through an adaptive sampling method, we constructed a lower dimensional effective model space, in which smoothing spline ANOVA models are estimated. We established the asymptotic convergence rate of smoothing splines via adaptive basis selection. More scalable computation make smoothing spline ANOVA models via adaptive basis selection an appealing method for ultra-large sample sequencing data. We demonstrated its excellent performance in both simulated studies and real examples.

4. PROPERTIES OF ADAPTIVE BASIS SAMPLING AND TECHNICAL PROOFS

4.1 Basic theoretical properties of adaptive basis sampling

This section presents some basic convergence properties of adaptive basis sampling to help gain some insights how it works. Since adaptive basis sampling involves the response variable, the standard argument for the asymptotic analysis of smoothing splines does not apply. The results in this section facilitate our study of asymptotic performance of the approximated smoothing spline estimator via adaptive basis sampling.

Consider the estimation of $\mathrm{E}\{\psi(X,Y)\}$ based on n i.i.d. observations $\{(x_i,y_i)\}_{i=1}^n$, where $\psi(x,y) \in \mathcal{L}^2(\mathcal{X},\mathcal{Y})$ is a generic multivariate function. The classical estimator is the sample average

$$\mathrm{E}_n(\psi) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i,y_i).$$

Suppose we use only a subsample by applying adaptive basis sampling. In the following, we shall study the asymptotic behavior of the subsample estimator. For simplicity in notation, we sometimes use either x_i or y_i to refer to (x_i,y_i) since the response and predictor variables come in pairs.

Adaptive basis sampling works as follows. First, we divide the range of $\{y_i\}_{i=1}^n$ into K slices. The number of observations in the k-th slice, $|S_k|$, is a random variable and it can be written as a sum of indicator functions, i.e. $|S_k| = \sum_{i=1}^n 1(y_i \in S_k)$. Next, n_k samples are drawn with replacement from the k-th slice. Then, we estimate $\mathrm{E}\{\psi(X,Y)\}$ using the aggregated subsamples $\{\psi(x_i^*,y_i^*)\}_{i=1}^{n^*} = \bigcup_{k=1}^K \{\psi(x_j^{*(k)},y_i^*)\}_{j=1}^{n_k}$,

where $n^* = \sum_{k=1}^{K} n_k$. The estimator is a weighted average and is written as

$$\mathbb{E}_n^*(\psi) = \sum_{k=1}^{K} \frac{|S_k|}{n} \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} \psi(x_j^{*(k)}, y_j^{*(k)}) \right\}. \tag{4.1}$$

Here, we use the operator notation $\mathbb{E}_n^*(\psi)$ to indicate that the sampling scheme works for a generic function ψ.

The linear operator $\mathbb{E}_n^*(\cdot)$ maps an element in $\mathcal{L}^2(\mathcal{X}, \mathcal{Y})$ to a random variable. Adaptive basis sampling implies that $\mathbb{E}_n^*(\psi)$ depends on the data $\{(x_i, y_i)\}_{i=1}^n$. In the following, we shall derive the conditional mean and variance of $\mathbb{E}_n^*(\psi)$ given the data and determine the magnitude of the distance of $\mathbb{E}_n^*(\psi)$ from $\mathrm{E}_n(\psi)$.

For each k, $1 \le k \le K$, $\{x_j^{*(k)}\}_{j=1}^{n_k}$ is a random draw from the k-th slice S_k. Thus, for $j = 1, \ldots, n_k$, the conditional mean of $\psi(x_j^{*(k)}, y_j^{*(k)})$ given the data is

$$\mathrm{E}\{\psi(x_j^{*(k)}, y_j^{*(k)}) | \{(x_i, y_i)\}_{i=1}^n\} = \frac{1}{|S_k|} \sum_{i=1}^{n} \psi(x_i, y_i) \mathbf{1}(y_i \in S_k). \tag{4.2}$$

It follows that the conditional mean of $\mathbb{E}_n^*(\psi)$ given the data is

$$\begin{aligned}
&\mathrm{E}\{\mathbb{E}_n^*(\psi) | \{(x_i, y_i)\}_{i=1}^n\} \\
&\quad = \mathrm{E}\left[\sum_{k=1}^{K} \frac{|S_k|}{n} \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} \psi(x_j^{*(k)}, y_j^{*(k)}) \right\} \middle| \{(x_i, y_i)\}_{i=1}^n \right] \\
&\quad = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \psi(x_i, y_i) \mathbf{1}(y_i \in S_k) = \frac{1}{n} \sum_{i=1}^{n} \psi(x_i, y_i) = \mathrm{E}_n(\psi).
\end{aligned}$$

Hence $\mathbb{E}_n^*(\psi)$ and $\mathrm{E}_n(\psi)$ have the same mean value, $\mathrm{E}(\psi)$.

In the k-th slice, for $j = 1, \ldots, n_k$, the conditional variance of $\psi(x_j^{*(k)}, y_j^{*(k)})$ given the data is bounded by its second order conditional moment whose explicit form can

be obtained by replacing ψ by ψ^2 in (4.2), i.e.

$$\begin{aligned}\mathrm{var}\{\psi(x_j^{*(k)}, y_j^{*(k)})|\{(x_i,y_i)\}_{i=1}^n\} &\leqslant \mathrm{E}\{\psi^2(x_j^{*(k)}, y_j^{*(k)})|\{(x_i,y_i)\}_{i=1}^n\} \\ &= \frac{1}{|S_k|}\sum_{i=1}^n \psi^2(x_i,y_i)\mathbf{1}(y_i \in S_k).\end{aligned} \tag{4.3}$$

Noticing that samples from the same slice and from different slices are mutually independent, we obtain that

$$\begin{aligned}&\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i,y_i)\}_{i=1}^n\} \\ &\quad= \mathrm{var}\left[\sum_{k=1}^K \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k}\psi(x_j^{*(k)}, y_j^{*(k)})\right\}\middle|\{(x_i,y_i)\}_{i=1}^n\right] \\ &\quad= \sum_{k=1}^K \frac{|S_k|^2}{n^2}\frac{1}{n_k}\mathrm{var}\{\psi(x_j^{*(k)}, y_j^{*(k)})|\{(x_i,y_i)\}_{i=1}^n\}.\end{aligned} \tag{4.4}$$

Lemma 4.1.1. *Suppose $n_k = n^*/K$, for $k = 1, \ldots, K$, then under the adaptive basis sampling scheme, the conditional variance of $\mathbb{E}_n^*(\psi)$ is bounded*

$$\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i,y_i)\}_{i=1}^n\} \leqslant \frac{K}{n^*}\,\frac{1}{n}\sum_{i=1}^n \psi^2(x_i,y_i) \tag{4.5}$$

and

$$\mathrm{E}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2 \leqslant \frac{K}{n^*}\,\mathrm{E}(\psi^2). \tag{4.6}$$

This lemma implies $\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)$ converges to zero in probability if $n^* \to \infty$ for ψ with $E\{\psi^2(X,Y)\} < \infty$. In other words, the subsample estimator, $\mathbb{E}_n^*(\psi)$, is a good surrogate of the usual estimator $\mathrm{E}_n(\psi)$.

Proof of Lemma 4.1.1 Since the variance of a random variable is bounded by its

second moment, (4.4) implies that

$$\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i,y_i)\}_{i=1}^n\} \leqslant \sum_{k=1}^K \frac{|S_k|}{n^2}\ \frac{1}{n_k}\sum_{i=1}^n \psi^2(x_i,y_i)\mathbf{1}(y_i \in S_k),$$

Applying (4.2) where ψ is replaced by ψ^2, we obtain that right-hand side of the above inequality equals

$$\sum_{k=1}^K \frac{|S_k|}{n^2}\ \frac{1}{n_k}\sum_{i=1}^n \psi^2(x_i,y_i)1(y_i \in S_k),$$

which in turn is upper bounded by

$$\sum_{k=1}^K \frac{1}{n}\frac{1}{n^*/K}\sum_{i=1}^n \psi^2(x_i,y_i)\mathbf{1}(y_i \in S_k) = \frac{K}{n^*}\ \frac{1}{n}\sum_{i=1}^n \psi^2(x_i,y_i)$$

with the fact that $n_k = n^*/K$ and $|S_k|/n \leqslant 1$. We thus have proved (4.5).

The condition mean of $\mathbb{E}_n^*(\psi)$ given the data has been proved to be $\mathrm{E}_n(\psi)$. Recall the definition of conditional variance, we have

$$\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i,y_i)\}_{i=1}^n\} = \mathrm{E}[\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2|\{(x_i,y_i)\}_{i=1}^n].$$

We obtain (4.6) immediately by taking expectation on both sides of the above, i.e.

$$\mathrm{E}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2 = \mathrm{E}[\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i,y_i)\}_{i=1}^n\}] \leqslant \frac{K}{n^*}\ \mathrm{E}(\psi^2).$$

4.2 Technical proofs

This section collects proofs of lemmas and theorems presented in Chapter 2 and Chapter 3. Since Chapter 2 can be seen as a special case of Chapter 3 in terms of technical proofs, we mainly focus on those in Chapter 3.

First, we present several ancillary lemmas.

4.2.1 Ancillary lemmas

We first present two lemmas in Gu (2013) that are useful for the proof of our main results.

Lemma 4.2.1. *Under Condition 3.2, as* $\lambda \to 0$*, one has*

$$\sum_{\nu} \frac{1}{1+\lambda\rho_\nu} = O(\lambda^{-1/r}).$$

This is part of Lemma 9.1 in Gu (2013).

Lemma 4.2.2. *Under Condition 3.1, 3.2 and 3.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$*,*

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i)h(x_i)w(\eta_0(x_i); y_i) = V(g,h) + o_p(\{(V+\lambda J)(g)(V+\lambda J)(h)\}^{1/2})$$

for all g *and* h *in* $\mathcal{H}$*.*

This is Lemma 9.16 in Gu (2013).

4.2.2 Proof of main results

We first present the proof of Lemma 3.3.1 and 3.3.2.

Proof of Lemma 3.3.1 According to the construction algorithm of the effective model space,

$$\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_j^*, \cdot),\ j = 1, \ldots, n^*\}.$$

For $h \in \mathcal{H} \ominus \mathcal{H}_E$, $h \perp g$ for $g \in \mathcal{H}_E$. Since $R_J(x_j^*, \cdot) \in \mathcal{H}_E$, we have $\langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}} = 0$ for $j = 1, \ldots, n^*$. On the other hand, $\mathcal{N}_J \subseteq \mathcal{H}_E$ implies $h \in \mathcal{H} \ominus \mathcal{N}_J = \mathcal{H}_J$. It then follows from the reproducing property of $R_J(\cdot,\cdot)$ on $\mathcal{H}_J$ that $h(x_j^*) = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}_J}$.

Noticing that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_J}$ can be obtained by restricting $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on

$\mathcal{H}_J$, we have

$$h(x_j^*) = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}_J} = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}} = 0,$$

for all $j = 1, \ldots, n^*$.

Proof of Lemma 3.3.2 By Lemma 3.3.1, given the selected samples $\{x_j^*\}_{j=1}^{n^*}$, for any $h \in \mathcal{H} \ominus \mathcal{H}_E$, we have

$$h(x_j^*) = 0 \qquad j = 1, \ldots, n^*.$$

Note that $\{x_j^*\}_{j=1}^{n^*}$ is the collection of $\{x_j^{*(k)}\}_{j=1}^{n_k}$ from $k = 1, \ldots, K$ slices, hence

$$\mathbb{E}_n^*\{h^2(X)w(\eta_0(X); Y)\} = \sum_{k=1}^{K} \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k} h^2(x_j^{*(k)})w(\eta_0(x_j^{*(k)}); y_j^{*(k)})\right\} = 0.$$

It follows that

$$V(h) = \int_{\mathcal{X}} h^2(x)v_{\eta_0}(x)f_X(x)\,dx = \mathrm{E}\{h(X)^2 v_{\eta_0}(X)\} - \mathbb{E}_n^*\{h^2(X)w(\eta_0(X); Y))\}. \tag{4.7}$$

By Condition 3.1, there exist a collection of functions $\phi_\nu \in \mathcal{H}$ and a sequence of nonnegative ρ_ν such that V and J are simultaneously diagonalized, i.e., $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu\,\delta_{\nu\mu}$. Use ϕ_ν's as basis functions and expand h as $h = \sum_\nu h_\nu\phi_\nu$, where $h_\nu = V(h, \phi_\nu)$. Then, (4.7) can be written as

$$V(h) = \mathrm{E}\left\{\left(\sum_\nu h_\nu\phi_\nu(X)\right)^2 v_{\eta_0}(X)\right\} - \mathbb{E}_n^*\left\{\left(\sum_\nu h_\nu\phi_\nu(X)\right)^2 w(\eta_0(X); Y)\right\}.$$

Due to the fact that $\mathrm{E}(\cdot)$ and $\mathbb{E}_n^*(\cdot)$ are both linear operators, we have

$$V(h) = \sum_{\nu}\sum_{\mu} h_\nu h_\mu \big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X)v_{\eta_0}(X)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}\big].$$

Applying the Cauchy-Schwarz inequality to obtain

$$V(h) \leqslant I^{1/2} \cdot \left\{\sum_{\nu}\sum_{\mu} h_\nu^2 h_\mu^2 (1+\lambda\rho_\nu)(1+\lambda\rho_\mu)\right\}^{1/2} \tag{4.8}$$

$$= I^{1/2} \cdot \sum_{\nu} h_\nu^2 (1+\lambda\rho_\nu) \tag{4.9}$$

where

$$I = \sum_{\nu}\sum_{\mu} \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X)v_{\eta_0}(X)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}\big]^2. \tag{4.10}$$

Since ϕ_ν's simultaneously diagonalize V and J,

$$\sum_{\nu} h_\nu^2(1+\lambda\rho_\nu) = (V+\lambda J)(h). \tag{4.11}$$

In light of (4.8), to bound $V(h)$, we need to investigate the magnitude of I whose expression is given in (4.10).

First, by inserting

$$\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} = \frac{1}{n}\sum_{i=1}^{n} \phi_\nu(x_i)\phi_\mu(x_i)w(\eta_0(x_i);y_i)$$

into the squared term in (4.10) and applying the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$, we

obtain

$$
\begin{aligned}
I \leqslant\; & 2 \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X) v_{\eta_0}(X)\} \\
& \qquad - \mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\} \big]^2 \\
& + 2 \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\} \\
& \qquad - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\big]^2 \\
\triangleq\; & 2I_1 + 2I_2.
\end{aligned}
$$

Next, we examine the magnitudes of I_1 and I_2 one by one.

Order of I_1.

Recall that $\mathrm{E}\{w(\eta_0(x); y)\} = v_{\eta_0}(x)$, then

$$
\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\}\big] = \mathrm{E}\{\phi_\nu(X)\phi_\mu(X) v_{\eta_0}(X)\}
$$

and

$$
\mathrm{var}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\}\big] = \frac{1}{n} \mathrm{var}\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\}.
$$

Therefore, the expectation of I_1 is

$$
\begin{aligned}
\mathrm{E}\, I_1 = & \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \mathrm{E}\big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X) v_{\eta_0}(X)\} \\
& \qquad - \mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\} \big]^2 \\
= & \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \frac{1}{n} \mathrm{var}\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\}.
\end{aligned}
$$

By Condition 3.4, $\mathrm{var}\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\} \leqslant c_3$ for some constant $c_3 < \infty$.

Hence, by Lemma 4.2.1,

$$\mathrm{E}\, I_1 \leqslant \frac{c_3}{n}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{-1}\lambda^{-2/r}). \tag{4.12}$$

Order of I_2.

The expectation of I_2 is

$$\mathrm{E}\, I_2 = \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\,\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\big]^2.$$

As in Lemma 4.1.1, we assume $n_k = n^*/K$ for all k and substitute $\psi(x,y)$ by $\phi_\nu(x)\phi_\mu(x)w(\eta_0(x);y)$ in (4.6) to obtain

$$\begin{aligned}
&\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\big]^2 \\
&\quad\leqslant \frac{K}{n^*}\mathrm{E}\{\phi_\nu^2(X)\phi_\mu^2(X)w^2(\eta_0(X);Y)\} \\
&\quad\leqslant \frac{K}{n^*}\,(c_3+1),
\end{aligned}$$

where the constant c_3 is the bound of $\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}$ in Condition 3.4. Again, by Lemma 4.2.1,

$$\mathrm{E}\, I_2 \leqslant \frac{K(c_3+1)}{n^*}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{*-1}\lambda^{-2/r}). \tag{4.13}$$

Putting (4.12) and (4.13) together and noticing $n^* \ll n$, we obtain

$$\mathrm{E}\, I \leqslant 2\ \mathrm{E}\, I_1 + 2\ \mathrm{E}\, I_2 = O(n^{*-1}\lambda^{-2/r}) + O(n^{-1}\lambda^{-2/r}) = O(n^{*-1}\lambda^{-2/r}).$$

Therefore $I = O_p(n^{*-1}\lambda^{-2/r})$ and $V(h) \leqslant (V + \lambda J)(h) \cdot O_p(n^{*-1/2}\lambda^{-1/r})$. The desired result follows from the fact $n^{*-1/2}\lambda^{-1/r} \to 0$.

Proof of Theorem 3.3.2 By the representer theorem, $\hat{\eta}$, the minimizer of (3.2) has an explicit form as in (3.3). Given the effective model space $\mathcal{H}_E$, let $\hat{\eta}_E$ be the projection of $\hat{\eta}$ to $\mathcal{H}_E$ relative to the reproducing kernel Hilbert space inner product. The proposed estimator $\hat{\eta}_A$ uses basis functions from $\mathcal{H}_E$ while $\hat{\eta}$ uses the full basis from $\mathcal{H}$.

According to Theorem 3.3.1, $\hat{\eta}$ converges to the true function η_0 with certain rate. Notice that

$$\hat{\eta}_A - \eta_0 = (\hat{\eta}_A - \hat{\eta}_E) + (\hat{\eta}_E - \hat{\eta}) + (\hat{\eta} - \eta_0).$$

It suffices to show that both $\hat{\eta}_E - \hat{\eta}$ and $\hat{\eta}_A - \hat{\eta}_E$ converge to zero at the same or a faster rate. We achieve this in two steps.

Step 1. We show that $\hat{\eta}_E$ converges to η_0 with the same rate as $\hat{\eta}$. To this end, note that $\hat{\eta} - \hat{\eta}_E \in \mathcal{H} \ominus \mathcal{H}_E \subseteq \mathcal{H}_J$ and $\hat{\eta} \in \mathcal{H}_E$, therefore $J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_E) = 0$.

For any functions $g, h \in \mathcal{H}$, define

$$A_{g,h}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} l\{(g + \alpha h)(x_i); y_i\} + \frac{\lambda}{2} J(g + \alpha h).$$

It can be easily shown that

$$\left.\frac{dA_{g,h}(\alpha)}{d\alpha}\right|_{\alpha=0} = \frac{1}{n}\sum_{i=1}^{n} u(g(x_i); y_i)h(x_i) + \lambda J(g, h). \tag{4.14}$$

Since $\hat{\eta}$ is the minimizer of (3.2) over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}$ and $h = \hat{\eta} - \hat{\eta}_E$. Thus, for this choice of g and h, the derivative in (4.14) is

zero. It follows that

$$\lambda J(\hat{\eta}, \hat{\eta} - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\}. \tag{4.15}$$

The fact that $J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_E) = 0$ implies $J(\hat{\eta} - \hat{\eta}_E)$ is equal to $J(\hat{\eta}, \hat{\eta} - \hat{\eta}_E)$. Thus

$$\lambda J(\hat{\eta} - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\} \triangleq S_1 + S_2, \tag{4.16}$$

where

$$S_1 = -\frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}(x_i); y_i) - u(\eta_0(x_i); y_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\},$$
$$S_2 = -\frac{1}{n}\sum_{i=1}^{n} u(\eta_0(x_i); y_i)\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\}.$$

We next study the orders of the two terms S_1 and S_2 under Conditions 3.1, 3.2, 3.4, and $\lambda \to 0$, $n\lambda^{2/r} \to \infty$.

For S_1, since $u(\eta(x), y)$ is differentiable with respect to $\eta(x)$, it follows by the mean value theorem and Condition 3.3 that there exists a constant $\gamma \in [c_1, c_2]$ such that

$$S_1 = -\frac{\gamma}{n}\sum_{i=1}^{n} w(\eta_0(x_i); y_i)\{\hat{\eta}(x_i) - \eta_0(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\}.$$

Applying Lemma 4.2.2 to the right hand side of the above, we have

$$\begin{aligned} |S_1| &= \gamma\, V(\hat{\eta} - \eta_0, \hat{\eta} - \hat{\eta}_E) + \{(V + \lambda J)(\hat{\eta} - \eta_0)(V + \lambda J)(\hat{\eta} - \hat{\eta}_E)\}^{1/2}\, o_p(1) \\ &= \{(V + \lambda J)(\hat{\eta} - \eta_0)(V + \lambda J)(\hat{\eta} - \hat{\eta}_E)\}^{1/2}\, O_p(1) \end{aligned}$$

For S_2, recall $\phi_\nu \in \mathcal{H}$ are eigenfunctions which simultaneously diagonalize V and J such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$. Write $\hat{\eta} - \hat{\eta}_E = \sum_\nu (\hat{\eta} - \hat{\eta}_E)_\nu \phi_\nu$,

where $(\hat{\eta} - \hat{\eta}_E)_\nu = V(\hat{\eta} - \hat{\eta}_E, \phi_\nu)$. Plugging it in S_2 and applying Cauchy-Schwarz inequality, we have

$$\begin{aligned} |S_2| &= \left| \sum_\nu (\hat{\eta} - \hat{\eta}_E)_\nu \left\{ \frac{1}{n} \sum_{i=1}^n u(\eta_0(x_i); y_i) \phi_\nu(x_i) \right\} \right| \\ &\leqslant \left\{ \sum_\nu \frac{\beta_\nu^2}{1 + \lambda \rho_\nu} \right\}^{1/2} \left\{ \sum_\nu (\hat{\eta} - \hat{\eta}_E)_\nu^2 (1 + \lambda \rho_\nu) \right\}^{1/2} \end{aligned}$$

where $\beta_\nu = \frac{1}{n} \sum_{i=1}^n u(\eta_0(x_i); y_i) \phi_\nu(x_i)$ possesses properties $\mathrm{E}(\beta_\nu) = 0$ and $\mathrm{var}(\beta_\nu) = \sigma^2/n$. In fact

$$\mathrm{E}(\beta_\nu) = \mathrm{E}\{u(\eta_0(X); Y) \phi_\nu(X)\} = \mathrm{E}_X \big[\mathrm{E}\{u(\eta_0(X); Y) | X\} \phi_\nu(X) \big] = 0$$

and

$$\begin{aligned} \mathrm{E}(\beta_\nu^2) &= \frac{1}{n} \mathrm{E}\{u^2(\eta_0(X); Y) \phi_\nu^2(X)\} = \frac{1}{n} \mathrm{E}_X \big[\mathrm{E}\{u^2(\eta_0(X); Y) | X\} \phi_\nu^2(X) \} \big] \\ &= \frac{\sigma^2}{n} \mathrm{E}_X \{ v_{\eta_0}(X) \phi_\nu^2(X) \} = \frac{\sigma^2}{n} V(\phi_\nu) = \frac{\sigma^2}{n}. \end{aligned}$$

Furthermore, by Lemma 4.2.1,

$$\mathrm{E} \left\{ \sum_\nu \frac{\beta_\nu^2}{1 + \lambda \rho_\nu} \right\} = \frac{\sigma^2}{n} \sum_\nu \frac{1}{1 + \lambda \rho_\nu} = O(n^{-1} \lambda^{-1/r}). \tag{4.17}$$

and it can be shown by a similar argument as in (4.11) that

$$\sum_\nu (\hat{\eta} - \hat{\eta}_E)_\nu^2 (1 + \lambda \rho_\nu) = (V + \lambda J)(\hat{\eta} - \hat{\eta}_E). \tag{4.18}$$

Combining (4.17) and (4.18), we obtain

$$S_2 \leqslant \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/(2r)}).$$

Now we are ready to determine the order of $(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)$. By Lemma 3.3.2, $V(\hat{\eta}-\hat{\eta}_E)$ is dominated by $\lambda J(\hat{\eta}-\hat{\eta}_E)$ since $\hat{\eta}-\hat{\eta}_E \in \mathcal{H}\ominus\mathcal{H}_E$. Thus, $(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)$ converges to zero at the same order as $\lambda J(\hat{\eta}-\hat{\eta}_E)$. Therefore, it follows (4.16) that

$$\begin{aligned}(V+\lambda J)(\hat{\eta}-\hat{\eta}_E) \asymp \lambda J(\hat{\eta}-\hat{\eta}_E) &= S_1 + S_2 \\ &\leqslant \{(V+\lambda J)(\hat{\eta}-\eta_0)(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)\}^{1/2} O_p(1) \\ &\quad + \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/(2r)}).\end{aligned}$$

After canceling out $\{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)\}^{1/2}$ and taking squares on both sides, we obtain

$$\begin{aligned}(V+\lambda J)(\hat{\eta}-\hat{\eta}_E) &\leqslant (V+\lambda J)(\hat{\eta}-\eta_0) O_p(1) + O_p(n^{-1}\lambda^{-1/r}) \\ &\asymp (V+\lambda J)(\hat{\eta}-\eta_0) \\ &= O_p(n^{-1}\lambda^{-1/r} + \lambda^p).\end{aligned}$$

Step 2. We show that $\hat{\eta}_A$, the smoothing spline estimator via adaptive sampling scheme, converges to η_0 with the same convergence rate as $\hat{\eta}_E$.

Since $\hat{\eta}$ is the minimizer of (3.2) over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}$ and $h = \hat{\eta}_A - \hat{\eta}_E$. Arguing as in the proof of (4.15), we have

$$\lambda J(\hat{\eta}, \hat{\eta}_A - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\}. \tag{4.19}$$

Since $\hat{\eta}_A$ is also the minimizer of (3.2) over $\mathcal{H}_E$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}_A$ and $h = \hat{\eta}_A - \hat{\eta}_E$. Thus, similar to the previous result, we have

$$\lambda J(\hat{\eta}_A, \hat{\eta}_A - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}_A(x_i); y_i)\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\}. \tag{4.20}$$

We subtract (4.19) from (4.20) to obtain

$$\lambda J(\hat{\eta}_A - \hat{\eta}, \hat{\eta}_A - \hat{\eta}_E) = \frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}(x_i); y_i) - u(\hat{\eta}_A(x_i); y_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\}.$$

Recall that $\hat{\eta}_E$ is the projection of $\hat{\eta}$ onto $\mathcal{H}_E$ and $\hat{\eta}_A - \hat{\eta}_E \in \mathcal{H}_E$, then $(\hat{\eta} - \hat{\eta}_E) \perp (\hat{\eta}_A - \hat{\eta}_E)$. Such orthogonality implies that $J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_A - \hat{\eta}_E) = 0$ and further

$$J(\hat{\eta}_A - \hat{\eta}_E) = J(\hat{\eta}_A - \hat{\eta}, \hat{\eta}_A - \hat{\eta}_E) + J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_A - \hat{\eta}_E) = J(\hat{\eta}_A - \hat{\eta}, \hat{\eta}_A - \hat{\eta}_E).$$

With this result, some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}_A(x_i); y_i) - u(\hat{\eta}_E(x_i); y_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\} + \lambda J(\hat{\eta}_A - \hat{\eta}_E) \tag{4.21}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}(x_i); y_i) - u(\hat{\eta}_E(x_i); y_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\} \tag{4.22}$$

By the mean value theorem, Condition 3.3 and Lemma 4.2.2, there exists a constant $\zeta \in [c_1, c_2]$ such that the left hand side of (4.21) equals

$$\zeta\, V(\hat{\eta}_A - \hat{\eta}_E) + o_p\{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_E)\} + \lambda J(\hat{\eta}_A - \hat{\eta}_E) = (V + \lambda J)(\hat{\eta}_A - \hat{\eta}_E)\{1 + o_p(1)\}.$$

Similarly the right hand side of (4.21) is bounded by

$$\{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\}^{1/2}O_p(1).$$

Combining the above two results, we obtain that

$$(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\{1+o_p(1)\}=\{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\}^{1/2}O_p(1).$$

Canceling out a term from both sides to obtain

$$(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\asymp(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)=O_p(n^{-1}\lambda^{-1/r}+\lambda^p). \tag{4.23}$$

Putting results from Step 1 and 2 together, we conclude the proof with the convergence rate

$$(V+\lambda J)(\hat{\eta}_A-\eta_0)=O_p(n^{-1}\lambda^{-1/r}+\lambda^p).$$

5. CONCLUSIONS

In this dissertation, we develop adaptive basis sampling for efficient computation of smoothing splines. The fast algorithm enable the classical nonparametric statistical model to deal with large data sets. We showed that, with a much smaller set of adaptively selected basis functions, the approximated smoothing splines can achieve the same rate of convergence of the smoothing splines with full basis. We demonstrate the excellent performance of the proposed smoothing spline estimator in both simulated studies and a deep Earth image analysis data set. Fast computation and asymptotic consistency make the smoothing splines via adaptive basis sampling an appealing method for large scale applications.

Motivated by the diverse types of data collected by next-generation sequencing technologies, we further focus on construct smoothing spline models for modeling counts data from exponential family distributions. Proper modeling of these data plays an important role in navigating the biological discovery. Two outstanding topics are carefully studied. First, we construct an smoothing spline ANOVA model to estimate gene expressions from RNA-seq data set. A joint modeling accounts for the well-known GC bias and reveals dynamics patterns over time. Further exploration such as isoform expression estimation can be implemented. The second example is genome-wide methylated region detection with bisulfite sequencing data. The fast computation of proposed method makes it possible to search over the whole genome and achieves an automatic procedure. In addition to real data examples, we also design simulation studies to demonstrated the excellent performance of our estimator.

The idea of adaptive basis sampling is innovative and is well suited with the

formulation of smoothing splines. Various models involving smoothing splines can borrow its strength in a similar manner, for example, mixed-effect models. Those extensions are future research topics.

REFERENCES

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005), "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, 122, 947–956.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009), "Asymptotic properties of penalized spline estimators," *Biometrika*, 96, 529–544.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., and Grimmond, S. M. (2008), "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, 5, 613–619.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008), "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning," *Nature*, 452, 215–219.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York: Wiley.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, 31, 377–403.

Dennis, J. E. and Schnabel, R. B. (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Philadelphia: SIAM, corrected reprint of the 1983 original.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012), "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, 485, 376–380.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008), "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, 36, e105.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Drevensek, S., Goussot, M., Duroc, Y., Christodoulidou, A., Steyaert, S., Schaefer, E., Duvernois, E., Grandjean, O., Vantard, M., Bouchez, D., and Pastuglia, M. (2012), "The Arabidopsis TRM1–TON1 interaction reveals a recruitment network common to plant cortical microtubule arrays and eukaryotic centrosomes," *Plant Cell*, 24, 178–191.

Duchon, J. (1977), "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables*, eds. Schemp, W. and Zeller, K., Berlin: Springer-Verlag, pp. 85–100.

Golub, G. and Van Loan, C. (1989), *Matrix Computations*, Baltimore, MD: The Johns Hopkins University Press, 2nd ed.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011), "The developmental transcriptome of Drosophila melanogaster," *Nature*, 471, 473–479.

Gu, C. (2013), *Smoothing Spline ANOVA Models*, vol. 297, New York: Springer, 2nd ed.

Gu, C. and Kim, Y.-J. (2002), "Penalized likelihood regression: General formulation and efficient approximation," *Canadian Journal of Statistics*, 30, 619–628.

Gu, C. and Qiu, C. (1994), "Penalized likelihood regression: a simple asymptotic

analysis," *Statist. Sin.*, 4, 297–304.

Gu, C. and Xiang, D. (2001), "Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited," *J. Comput. Graph. Statist.*, 10, 581–591.

Hansen, K. D., Langmead, B., Irizarry, R. A., et al. (2012), "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions," *Genome Biology*, 13, R83.

Hastie, T. J. (1996), "Pseudosplines," *J. Roy. Statist. Soc. Ser. B*, 58, 379–396.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012), "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International Journal of Epidemiology*, 41, 200–209.

Ji, L., Sasaki, T., Sun, X., Ma, P., Lewis, Z. A., and Schmitz, R. J. (2014), "Methylated DNA is over-represented in whole-genome bisulfite sequencing data," *Frontiers in Genetics*, 5, 341.

Jiang, H. and Wong, W. H. (2009), "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, 25, 1026–1032.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007), "Genome-wide mapping of in vivo protein-DNA interactions ," *Science*, 316, 1497–502.

Kim, Y.-J. and Gu, C. (2004), "Smoothing spline Gaussian regression: More scalable computation via efficient approximation," *J. Roy. Statist. Soc. Ser. B*, 66, 337–356.

Kuan, P. F., Chung, D., Pan, G., Thomson, J. A., Stewart, R., and Keles, S. (2011), "A statistical framework for the analysis of ChIP-Seq data," *Journal of the American Statistical Association*, 106, 891–903.

Leung, Y. F., Ma, P., Link, B. A., and Dowling, J. E. (2008), "Factorial microarray analysis of zebrafish retinal development," *Proceedings of the National Academy of Sciences*, 105, 12909–12914.

Li, J., Jiang, H., and Wong, W. H. (2010), "Modeling non-uniformity in short-read rates in RNA-Seq data ," *Genome Biology*, 11, R50.

Li, K. C. (1991), "Sliced inverse regression for dimension reduction," *J. Amer. Statist. Assoc.*, 86, 316–327.

Lindeboom, J. J., Nakamura, M., Hibbel, A., Shundyak, K., Gutierrez, R., Ketelaar, T., Emons, A. M. C., Mulder, B. M., Kirik, V., and Ehrhardt, D. W. (2013), "A mechanism for reorientation of cortical microtubule arrays driven by microtubule severing," *Science*, 342, 1245533.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008), "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis," *Cell*, 133, 523–536.

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *Journal of Machine Learning Research*, 10, 2295–2328.

Liu, Z. and Guo, W. (2010), "Data driven adaptive spline smoothing," *Statist. Sin.*, 20, 1143–1163.

Luo, Z. and Wahba, G. (1997), "Hybrid adaptive splines," *J. Amer. Statist. Assoc.*, 92, 107–116.

Ma, P., Wang, P., Tenorio, L., de Hoop, M. V., and van der Hilst, R. D. (2007), "Imaging of structure at and near the core mantle boundary using a generalized Radon transform: 2. Statistical inference of singularities," *J. Geophys. Res.*, 112, B08303.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008), "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, 5, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and

Snyder, M. (2008), "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, 320, 1344–1349.

Nychka, D. (1988), "Bayesian confidence intervals for smoothing splines," *J. Amer. Statist. Assoc.*, 83, 1134–1143.

Pintore, A., Speckman, P., and Holmes, C. C. (2006), "Spatially adaptive smoothing splines," *Biometrika*, 93, 113–125.

Reinsch, C. H. (1967), "Smoothing by spline functions," *Numerische Mathematik*, 10, 177–183.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric regression*, Cambridge, UK: Cambridge University Press.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

Silverman, B. W. (1982), "On the estimation of a probability density function by the maximum penalized likelihood method," *Ann. Statist.*, 10, 795–810.

Stone, C. J. (1982), "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, 10, 1040–1053.

Sun, X., Dalpiaz, D., Wu, D., Liu, J. S., and Ma, P. (2015), "Statistical modeling of time course RNA-Seq via negative binomial mixed model," *Bioinformatics*, 0, 000–000.

Tenorio, L., Andersson, F., de Hoop, M., and Ma, P. (2011), "Data analysis tools for uncertainty quantification of inverse problems," *Inverse Problems*, 27, 045001.

Utreras, F. (1981), "Optimal smoothing of noisy data using spline functions," *SIAM J. Sci. Statist. Comput.*, 2, 349–362.

van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S. H., Ma, P., and Tenorio, L. (2007), "Seismo-stratigraphy and thermal structure of Earth's core-mantle boundary region," *Science*, 315, 1813–1817.

Wahba, G. (1983), "Bayesian "confidence intervals" for the cross-validated smoothing spline," *J. Roy. Statist. Soc. Ser. B*, 45, 133–150.

— (1985), "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *Ann. Statist.*, 13, 1378–1402.

— (1990), *Spline Models for Observational Data*, vol. 59 of CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy : the 1994 Neyman Memorial Lecture," *Ann. Statist.*, 23, 1865–1895.

Wang, P., de Hoop, M. V., van der Hilst, R. D., Ma, P., and Tenorio, L. (2006), "Imaging of structure at and near the core mantle boundary using a generalized Radon transform: 1. Construction of image gathers," *J. Geophys. Res.*, 111, B12304.

Wang, X., Du, P., and Shen, J. (2013), "Smoothing splines with varying smoothing parameter," *Biometrika*, 100, 955–970.

Wang, X., Shen, J., Ruppert, D., et al. (2011), "On the asymptotics of penalized spline smoothing," *Electronic Journal of Statistics*, 5, 1–17.

Wang, Y. (2011), *Smoothing Splines: Methods and Applications*, Boca Raton: Chapman and Hall.

Weinberger, H. F. (1974), *Variational Methods for Eigenvalue Approximation*, Philadelphia: SIAM.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bahler, J. (2008), "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution," *Nature*, 453, 1239–U39.

Xiao, L., Li, Y., and Ruppert, D. (2013), "Fast bivariate P-splines: the sandwich

smoother," *J. Roy. Statist. Soc. Ser. B*, 75, 577–599.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., and Klein, B. (2004), "Variable Selection and Model Building Via Likelihood Basis Pursuit," *J. Amer. Statist. Assoc.*, 99, 659–672.

Zheng, W., Chung, L. M., and Zhao, H. (2011), "Bias detection and correction in RNA-Sequencing data," *BMC Bioinformatics*, 12, 290.